Large Dimensional Analysis of Robust M-Estimators of Covariance With Outliers

David Morales-Jimenez, Romain Couillet, Matthew Mckay

Large Dimensional Analysis of Robust M-Estimators of Covariance with Outliers

David Morales-Jimenez*, Romain Couillet†, Matthew R. McKay*

***Abstract*—A large dimensional characterization of robust M-estimators of covariance (or scatter) is provided under the assumption that the dataset comprises independent (essentially Gaussian) legitimate samples as well as arbitrary deterministic samples, referred to as outliers. Building upon recent random matrix advances in the area of robust statistics, we specifically show that the so-called Maronna M-estimator of scatter asymptotically behaves similar to well-known random matrices when the population and sample sizes grow together to infinity. The introduction of outliers leads the robust estimator to behave asymptotically as the weighted sum of the sample outer products, with a constant weight for all legitimate samples and different weights for the outliers. A fine analysis of this structure reveals importantly that the propensity of the M-estimator to attenuate (or enhance) the impact of outliers is mostly dictated by the alignment of the outliers with the inverse population covariance matrix of the legitimate samples. Thus, robust M-estimators can bring substantial benefits over more simplistic estimators such as the per-sample normalized version of the sample covariance matrix, which is not capable of differentiating the outlying samples. The analysis shows that, within the class of Maronna's estimators of scatter, the Huber estimator is most favorable for rejecting outliers. On the contrary, estimators more similar to Tyler's scale invariant estimator (often preferred in the literature) run the risk of inadvertently enhancing some outliers.**

***Index Terms*—Robust statistics, M-estimation, outliers.**

I. Introduction

The growing momentum of big data applications along with the recent advances in large dimensional random matrix theory have raised much interest for problems in statistics and signal processing under the assumption of large but similar population dimension N and sample size n. Due to the intrinsic complexity of large dimensional random matrix theory, as compared to classical statistics where N is fixed and $n \to \infty$, most of the classical applications were concerned with sample covariance matrix (SCM) based methods (as in e.g., [1, 2] for source detection or [3] for subspace estimation). Only recently have other random matrix structures started to be explored which are adequate to deal with more advanced statistical problems; see for instance [4] on Toeplitz random matrix structures, or [5] on kernel random matrices. Of particular interest is the structure of robust M-estimators of covariance (or scatter), which have very recently come to a better understanding in the large dimensional regime and is the focus of the present work.

The field of robust M-estimation, born with the early works of Huber [6], roughly consists in improving classical Gaussian maximum-likelihood estimators, such as the sample mean or SCM, into estimators that (unlike the classical estimators) are resilient to both the possibly heavy-tailed nature of the observed data or the presence of outliers in the dataset. Assuming observation data of known zero mean, robust estimators of the population covariance matrix, referred to as robust M-estimators of scatter, were proposed successively in [6] for data composed of a majority of independent Gaussian samples and a few outliers and then in [7] and [8] for elliptically distributed or arbitrary scaled Gaussian data.

But the analysis for each given N, n of the aforementioned robust estimators of scatter, which often take the form of solutions of implicit equations, is in general intractable. In a series of recent works [9–12] (see also [13, 14] for applications), this limitation was alleviated by considering the random matrix regime where both N, n are large and commensurable. These works have shown that in this regime several classes of robust estimators of scatter (Maronna, Tyler, and regularized Tyler) behave similar to simpler and explicit random matrix models, which are fully understandable via (now standard) random matrix methods. Nonetheless, all these works were pursued under the assumption that the input data are independent and follow a zero-mean elliptical distribution. One of the salient outcomes of these works is that, under elliptical inputs, the Tyler and regularized Tyler estimators asymptotically behave similar to the SCM of the normalized data,[1] henceforth referred to as the normalized SCM, and therefore do not provide any apparent gain in robustness versus simpler sample covariance estimators.

This fact, however, fundamentally disregards the important role of robust estimators as arbitrary outlier rejectors. In the present work, we shall consider data comprising both legitimate data (that are essentially independent Gaussian samples) and a certain (a priori unknown) amount of arbitrary deterministic outliers. Focusing our attention specifically to the (larger) class of Maronna's M-estimators of scatter, similar to all of the aforementioned works and following the approach in [9], we will show that in this setting the robust estimator of scatter behaves similar for large N, n to an explicit and easily understood random matrix. But it will appear, unlike in

*D. Morales-Jimenez and M. R. McKay are with Dept. Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon (Hong Kong). (e-mail:{eedmorales,eemckay}@ust.hk)

†R. Couillet is with CNRS-CentraleSupélec-Université Paris-Sud, 91192 Gif-sur-Yvette, France (romain.couillet@centralesupelec.fr).

The work of D. Morales-Jimenez and M. R. McKay was supported by the Hong Kong Research Grants Council under grant number 16206914. Couillet's work is supported by the ERC MORE EC–120133.

[1]This being valid up to second-order fluctuations [11].

[9–12], that this random matrix no longer behaves similar to the normalized SCM. Our main finding is that, under suitable conditions, the robust estimator of scatter manages to attenuate (to some extent) the impact of the deterministic outliers, which the SCM (or normalized SCM) may not be capable of. Calling $\mathbf{C}_N$ the population covariance matrix of the legitimate data, $\mathbf{a}_i \in \mathbb{C}^N$ the i-th outlier, and assuming the number of outliers is small compared to n, it will be demonstrated that the rejection power of the robust estimator of scatter is monotonically related to the quadratic form $\mathbf{a}_i^\dagger \mathbf{C}_N^{-1} \mathbf{a}_i$. This shows that, if $\mathbf{C}_N$ is (invertible but) essentially of low rank, $\mathbf{a}_i^\dagger \mathbf{C}_N^{-1} \mathbf{a}_i$ can take large values and thus $\mathbf{a}_i$ is likely to be suppressed. If $\mathbf{a}_i^\dagger \mathbf{C}_N^{-1} \mathbf{a}_i$ is quite small instead, an inverse effect of outlier enhancement may appear that needs be controlled by an appropriate choice of estimator within Maronna's class. We shall show that such an estimator should resemble the original Huber estimator from [6] and substantially differ from the Tyler estimator.

In the remainder of the article, we provide a rigorous statement of our main results. The problem at hand is discussed in Section II and our main results introduced in Section III, all proofs being deferred to the appendices. Special attention will then be made on the analytically tractable cases where the number of outliers is either small (Section IV) or random i.i.d. (Section V). Concluding remarks are provided in Section VI.

Notations: The superscript $(\cdot)^\dagger$ stands for Hermitian transpose in the complex case or transpose in the real case. The norm $\|\cdot\|$ is the spectral norm for matrices and the Euclidean norm for vectors. The Dirac measure at point x is denoted δ_x and $\mathbf{1}_A$ stands for the indicator function with A the corresponding inclusion event. The imaginary unit is denoted $\imath = \sqrt{-1}$ and $\Im[\cdot]$ stands for the imaginary part. The set $\mathbb{R}^+$ is defined as $\{x : x \geq 0\}$ and $\mathbb{C}^+ = \{z \in \mathbb{C},\ \Im[z] > 0\}$. The support of a distribution function F is denoted by $\mathrm{Supp}(F)$. The ordered eigenvalues of a Hermitian (or symmetric) matrix $\mathbf{X}$ of size $N \times N$ are denoted $\lambda_1(\mathbf{X}) \leq \ldots \leq \lambda_N(\mathbf{X})$. For $\mathbf{A}, \mathbf{B}$ Hermitian, $\mathbf{A} \succ \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive definite. The notation $\mathrm{diag}(\mathbf{X})$ stands for the diagonal matrix composed of the diagonal elements of matrix $\mathbf{X}$ and $\mathrm{diag}(\mathbf{x})$ the diagonal matrix composed of the elements of vector $\mathbf{x}$ on the diagonal. The arrow $\xrightarrow{\text{a.s.}}$ designates almost sure convergence and $\Rightarrow$ stands for weak convergence.

II. System Model and Motivation

For $\varepsilon_n \in \mathbb{R}$ such that $n\varepsilon_n \in \{1, \ldots, n\}$, let

$$\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_{(1-\varepsilon_n)n}, \mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n}\right] \in \mathbb{C}^{N \times n}$$

where $\mathbf{y}_i = \mathbf{C}_N^{1/2} \mathbf{x}_i \in \mathbb{C}^N$, $i = 1, \ldots, (1-\varepsilon_n)n$, are independent across i, $\mathbf{C}_N \in \mathbb{C}^{N \times N}$ is deterministic Hermitian positive definite, and $\mathbf{x}_i$ has zero mean, unit variance and finite $(8+\eta)$-th order moment entries for some $\eta > 0$, while $\mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n} \in \mathbb{C}^N$ are arbitrary deterministic vectors.[2] We shall further assume that, as $N \to \infty$, $\limsup_N \|\mathbf{C}_N\| < \infty$.

The vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{(1-\varepsilon_n)n}$ will be considered the legitimate data, while $\mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n}$ are deterministic unknown outliers. It is important to note at this point that all estimators of $\mathbf{C}_N$ considered in the following are invariant to column permutations in $\mathbf{Y}$ so that we can freely assume the first columns of $\mathbf{Y}$ to be the legitimate data and the last columns to be the outliers. Note also that we consider here a more general setting than Gaussian legitimate data as we merely request the $\mathbf{x}_i$'s to have independent normalized entries with some bounded moment condition.

Although $\mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n}$ are arbitrary, for technical reasons we shall need the following control.

Assumption 1. $\limsup_n \| \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} \mathbf{C}_N^{-1/2} \mathbf{a}_i \mathbf{a}_i^\dagger \mathbf{C}_N^{-1/2} \| < \infty$.

Note that, if $\limsup_n \varepsilon_n n < \infty$, Assumption 1 reduces to $\limsup_n \max_{1 \leq i \leq \varepsilon_n n} \frac{1}{N} \mathbf{a}_i^* \mathbf{C}_N^{-1} \mathbf{a}_i < \infty$.

If one were aware of the presence and position of outliers in the dataset, then the natural estimator for $\mathbf{C}_N$ (up to renormalization by $1 - \varepsilon_n$) would read $\frac{1}{n} \mathbf{Y}^{\mathrm{o}} \mathbf{Y}^{\mathrm{o}\dagger}$ with $\mathbf{Y}^{\mathrm{o}} = [\mathbf{y}_1, \ldots, \mathbf{y}_{(1-\varepsilon_n)n}]$; this estimator, which we shall refer to as the Oracle estimator (hence the "o" superscript), merely consists in a SCM with discarded outliers. For lack of knowing the outliers presence and positions, the immediate alternative estimate for $\mathbf{C}_N$ is the SCM, which reads here $\frac{1}{n} \mathbf{Y} \mathbf{Y}^\dagger$. If one is only interested in estimating any scaled version of $\mathbf{C}_N$, then, to mitigate the negative impact of outliers with arbitrarily large norm, a simple robust procedure consists in estimating $\mathbf{C}_N$ via the normalized SCM $\frac{1}{n} \mathbf{Y}^{\mathrm{n}} \mathbf{Y}^{\mathrm{n}\dagger}$, where $\mathbf{Y}^{\mathrm{n}} \triangleq \mathbf{Y} \mathrm{diag}(\frac{1}{N} \mathbf{Y}^\dagger \mathbf{Y})^{-\frac{1}{2}}$. This matrix has the advantage of avoiding arbitrarily large biases in the estimation of $\mathbf{C}_N$. However, being only based on a per-data norm control, $\frac{1}{n} \mathbf{Y}^{\mathrm{n}} \mathbf{Y}^{\mathrm{n}\dagger}$ does not take into account the fact that outliers can also be detected if they significantly differ, not just in norm, from the majority of the data. The robust estimators of scatter, introduced by Huber [6] and later studied by Maronna [7], were precisely designed for this purpose. Our objective here is to finely understand this outlier identification and mitigation procedure by means of a large random matrix analysis.

To be able to define a robust M-estimator of scatter in the sense of Maronna under the presence of arbitrary outlier vectors, a constraint must be set on ε_n and N. In particular, as n grows large, we shall require that $n(1-\varepsilon_n)/N$ (and not only n/N) be always beyond one.

Assumption 2 (Growth rate)**.** *As $n \to \infty$ $\varepsilon_n \to \varepsilon \in [0, 1)$ and $c_n \triangleq \frac{N}{n} \to c$ with $0 < c < 1 - \varepsilon$.*

We then define Maronna's M-estimator of scatter $\hat{\mathbf{C}}_N$ as a solution, when it exists, to the equation in $\mathbf{Z}$

$$\mathbf{Z} = \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} u\left(\frac{1}{N} \mathbf{y}_i^\dagger \mathbf{Z}^{-1} \mathbf{y}_i\right) \mathbf{y}_i \mathbf{y}_i^\dagger + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} u\left(\frac{1}{N} \mathbf{a}_i^\dagger \mathbf{Z}^{-1} \mathbf{a}_i\right) \mathbf{a}_i \mathbf{a}_i^\dagger. \quad (1)$$

where $u : [0, \infty) \to (0, \infty)$ is continuous, non-increasing, and such that $\phi(x) \triangleq xu(x)$ is increasing with $\lim_{x \to \infty} \phi(x) \triangleq \phi_\infty$ and $(1-\varepsilon)^{-1} < \phi_\infty < c^{-1}$. Note that the latter assumption on ϕ_∞ is equivalent to that in [9] with a slight modification accounting for the presence of outliers.

[2] As shall be seen in Section V, the vectors $\mathbf{a}_i$'s can be considered random as long as they are independent of the $\mathbf{y}_i$'s.

A standard choice for the function u is $u = u_{\rm S}$, where, for some $t > 0$,

$$u_{\rm S}(x) \triangleq \frac{1+t}{t+x} \tag{2}$$

which, for an appropriate t, turns $\hat{\mathbf{C}}_N$ into the maximum-likelihood estimator of $\mathbf{C}_N$ when the columns of $\mathbf{Y}$ are independent multivariate Student vectors (hence the superscript "S"). As $t \to 0$, $\hat{\mathbf{C}}_N$ converges to one version of the so-called Tyler estimator [8], as shown in [15].[3] We shall however restrict our study here to Maronna's class of estimators. Of particular interest in the present work is another function u, which we shall (somewhat abusively[4]) refer to as Huber's estimator function $u_{\rm H}$, defined, for some $t > 0$, as

$$u_{\rm H}(x) \triangleq \max\left\{1, \frac{1+t}{t+x}\right\}. \tag{3}$$

This function has the particularity of being constant for all $x \leq 1$, which will be later seen as an important property.

III. Main Result

From the problem setting, Assumption 2, and [16, Thm. 2.3], it is easily seen that, with probability one, the solution of (1) is unique for all large n and thus $\hat{\mathbf{C}}_N$ is unequivocally defined. In the same spirit as in [9, 10] (and with similar notations), our first objective is to find an explicit tight approximation of the implicitly defined $\hat{\mathbf{C}}_N$ in the regime where $N, n \to \infty$ as per Assumption 2. Our main result unfolds as follows.

Theorem 1 (Asymptotic Behavior)**.** *Let Assumptions 1–2 hold and let $\hat{\mathbf{C}}_N$ the solution to* (1) *(unique for all large n, with probability one). Then, as $n \to \infty$,*

$$\left\| \hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N \right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{\mathbf{S}}_N \triangleq v\left(\gamma_n\right) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} \mathbf{y}_i \mathbf{y}_i^\dagger + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v\left(\alpha_{i,n}\right) \mathbf{a}_i \mathbf{a}_i^\dagger$$

with $v(x) = u\left(g^{-1}(x)\right)$, $g(x) = x/(1 - c\phi(x))$, and $(\gamma_n, \alpha_{1,n}, \ldots, \alpha_{\varepsilon_n n, n})$ the solution to

$$\begin{aligned}
\gamma_n &= \frac{1}{N} \operatorname{tr} \mathbf{C}_N \left(\frac{(1-\varepsilon) v(\gamma_n)}{1 + c v(\gamma_n) \gamma_n} \mathbf{C}_N + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v(\alpha_{j,n}) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \\
\alpha_{i,n} &= \frac{1}{N} \mathbf{a}_i^\dagger \left(\frac{(1-\varepsilon) v(\gamma_n)}{1 + c v(\gamma_n) \gamma_n} \mathbf{C}_N + \frac{1}{n} \sum_{j \neq i} v(\alpha_{j,n}) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{a}_i
\end{aligned} \tag{4}$$

for $i = 1, \ldots, \varepsilon_n n$. In particular, from [17, Thm. 4.3.7],

$$\max_{1 \leq i \leq N} \left| \lambda_i(\hat{\mathbf{C}}_N) - \lambda_i(\hat{\mathbf{S}}_N) \right| \xrightarrow{\text{a.s.}} 0.$$

[3] As opposed to Maronna's class of estimators, Tyler estimator is only defined up to a constant factor; thus it estimates $\mathbf{C}_N$ up to a scale parameter.

[4] Huber's original estimator takes the form $u(x) = \max\{\alpha, \beta/x\}$ for some α, β, hence with additional parameters and with $t = 0$. However, uniqueness of $\hat{\mathbf{C}}_N$ is not guaranteed for $t = 0$ and, in the random matrix limit, $\alpha = \beta = 1$ is a particularly appealing choice.

Remark 1 (Function v)**.** *The function v defined in Theorem 1 was already introduced in [9] and uses, through g, the assumption that $\phi(x) < c^{-1}$. It has essentially the same general properties as u in that it is continuous, non-increasing and such that $\psi(x) \triangleq xv(x)$ is increasing and bounded with $\lim_{x\to\infty} \psi(x) \triangleq \psi_\infty = \phi_\infty/(1 - c\phi_\infty)$.*

Remark 2 (Relation to previous results)**.** *Taking $\varepsilon_n = 0$, Theorem 1 reduces to the result obtained in [18] and [12], i.e., $\hat{\mathbf{S}}_N = v(\gamma_n)\frac{1}{n}\mathbf{Y}\mathbf{Y}^\dagger$. In this case,* (4) *reduces to*

$$\gamma_n = \frac{1 + c v(\gamma_n) \gamma_n}{v(\gamma_n)}$$

which, after basic algebra, entails $\gamma_n = \phi^{-1}(1)/(1-c)$ and $v(\gamma_n) = 1/\phi^{-1}(1)$.

Theorem 1 allows us to transfer many properties of the implicit matrix $\hat{\mathbf{C}}_N$ into the more tractable matrix $\hat{\mathbf{S}}_N$, the random matrix structure of which is well known and has been studied as early as in [19]. The structure of $\hat{\mathbf{S}}_N$ is particularly interesting as it mostly consists of two terms: the sum of outer products of the legitimate data scaled by a constant factor $v(\gamma_n)$ along with a per-sample weighted sum of the outer products of the outlying data. Therefore, as one would expect, $\hat{\mathbf{C}}_N$ sets a specific emphasis (either small or large) on each outlying sample while maintaining all legitimate data under constant weight. We expect here that, as opposed to the SCM that provides no control on the data or to the normalized SCM that merely normalizes the outliers, $\hat{\mathbf{C}}_N$ will appropriately ensure a reduction of the outlier impact by letting $v(\alpha_{j,n})$ be quite small compared with $v(\gamma_n)$, especially if ε_n is small.

An immediate corollary of Theorem 1 concerns the large N eigenvalue distribution of $\hat{\mathbf{C}}_N$ and reads as follows.

Corollary 1 (Spectral Distribution)**.** *Define the empirical spectral distribution $F_N^{\hat{\mathbf{C}}_N}(x) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{\lambda_i(\hat{\mathbf{C}}_N) \leq x\}}$ for $x \in \mathbb{R}$. Then, under the setting of Theorem 1,*

$$F_N^{\hat{\mathbf{C}}_N}(x) - F_N(x) \Rightarrow 0$$

almost surely as $n \to \infty$, where $F_N(x)$ is a real distribution function with density defined via its Stieltjes transform $m_N(z)$ (i.e.,[5] $m_N(z) \triangleq \int (t-z)^{-1} dF_N(t)$) given for all $z \in \mathbb{C}^+$ by

$$m_N(z) = \frac{1}{N} \operatorname{tr} \left(\frac{(1-\varepsilon) v(\gamma_n)}{1 + e_N(z)} \mathbf{C}_N + \mathbf{A}_N - z \mathbf{I}_N \right)^{-1}$$

with $\mathbf{A}_N \triangleq \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v\left(\alpha_{i,n}\right) \mathbf{a}_i \mathbf{a}_i^\dagger$ and $e_N(z)$ the unique solution in $\mathbb{C}^+$ of the equation

$$e_N(z) = \frac{v(\gamma_n)}{n} \operatorname{tr} \mathbf{C}_N \left(\frac{(1-\varepsilon) v(\gamma_n)}{1 + e_N(z)} \mathbf{C}_N + \mathbf{A}_N - z \mathbf{I}_N \right)^{-1}.$$

In the appendix, it is importantly shown that $\limsup_N \|\hat{\mathbf{C}}_N\| < \infty$ a.s. (as a result of $\limsup_N \|\hat{\mathbf{S}}_N\| < \infty$

[5] Recall that any distribution function F is uniquely defined by its Stieltjes transform $m(z)$ by the fact that, for all continuity points a, b of F,

$$F(b) - F(a) = \lim_{y \downarrow 0} \int_a^b \Im[m(t + \imath y)] dt.$$

a.s.). This implies that $F_N^{\hat{\mathbf{C}}_N}$ and F_N have compact supports and are fully determined by their respective moments $M_{N,k}^{\hat{\mathbf{C}}_N} \triangleq \int t^k dF_N^{\hat{\mathbf{C}}_N}(t)$ and $M_{N,k} \triangleq \int t^k dF_N(t)$, $k = 1, 2, \ldots$, which satisfy $M_{N,k}^{\hat{\mathbf{C}}_N} - M_{N,k} \xrightarrow{\text{a.s.}} 0$ (by the dominated convergence theorem). While F_N is defined via its deterministic but implicit Stieltjes transform, the $M_{N,k}$ can be retrieved explicitly using successive derivatives of the moment generating formula (for $|z| < 1/\sup(\text{Supp}(F_N))$)

$$m_N(1/z) = -\sum_{k=0}^{\infty} z^{k+1} M_{N,k}.$$

Precisely, we obtain here the following result.

Corollary 2 (Moments)**.** *For F_N defined in Corollary 1, letting $M_{N,p} \triangleq \int t^p dF_N(t)$, $p = 1, 2, \ldots$,*

$$M_{N,p} = \frac{(-1)^p}{p!} \frac{1}{N} \operatorname{tr} \mathbf{T}_p$$

where $\mathbf{T}_p$ is obtained from the following recursive formulas

$$\begin{aligned}
\mathbf{T}_{p+1} &= -\sum_{i=0}^{p} \mathbf{T}_{p-i} \mathbf{A}_N \mathbf{T}_i + \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} \mathbf{T}_{p-i} \mathbf{Q}_{i-j+1} \mathbf{T}_j \\
\mathbf{Q}_{p+1} &= (p+1) f_p (1-\varepsilon) v(\gamma_n) \mathbf{C}_N \\
f_{p+1} &= \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} (p-i+1) f_j f_{i-j} \beta_{p-i} \\
\beta_{p+1} &= v(\gamma_n) \frac{1}{n} \operatorname{tr} \mathbf{C}_N \mathbf{T}_{p+1},
\end{aligned}$$

with initial values $\mathbf{T}_0 = \mathbf{I}_N$, $f_0 = -1$, $\beta_0 = v(\gamma_n) \frac{1}{n} \operatorname{tr} \mathbf{C}_N$. In particular,

$$\begin{aligned}
M_{N,1} &= \frac{1}{N} \operatorname{tr} \left[\mathbf{A}_N + (1-\varepsilon) v(\gamma_n) \mathbf{C}_N \right] \\
M_{N,2} &= \frac{1}{N} \operatorname{tr} \Big[\mathbf{A}_N^2 + 2(1-\varepsilon) v(\gamma_n) \mathbf{C}_N \mathbf{A}_N \\
&\quad + (1-\varepsilon)^2 v^2(\gamma_n) \mathbf{C}_N^2 + \left[\frac{1}{n} \operatorname{tr} \mathbf{C}_N \right] (1-\varepsilon) v^2(\gamma_n) \mathbf{C}_N \Big].
\end{aligned}$$

Albeit having characterized the random matrix $\hat{\mathbf{S}}_N$, which approximates the behavior of $\hat{\mathbf{C}}_N$ for large N, n, it is quite challenging to gain a good intuitive understanding of the weight structure as the expression (4) relating γ_n to the $\alpha_{i,n}$'s is still implicit (while being deterministic). To get more insight on the properties of $\hat{\mathbf{C}}_N$, we shall successively consider two specific scenarios that simplify the system (4).

IV. Finitely Many Outliers Scenario

Let us first assume that $\varepsilon_n n = K$ is maintained constant as $n \to \infty$ (thus $\varepsilon = 0$). Recall that, in this scenario, Assumption 1 can be replaced by the sufficient condition $\limsup_N \max_{1 \le i \le \varepsilon_n n} \frac{1}{N} \mathbf{a}_i^* \mathbf{C}_N^{-1} \mathbf{a}_i < \infty$. In the appendix, it is shown that γ_n cannot grow unbounded with n. As such, by a rank-one perturbation argument iterated K times, see e.g., [19, Lemma 2.6], we find that

$$\gamma_n - \frac{1 + c v(\gamma_n) \gamma_n}{v(\gamma_n)} = \mathcal{O}(1/N)$$

which ensures by Remark 2 that

$$\gamma_n = \frac{\phi^{-1}(1)}{1-c} + \mathcal{O}(1/N).$$

We shall denote next $\gamma \triangleq \frac{\phi^{-1}(1)}{1-c}$ (and thus $v(\gamma) = 1/\phi^{-1}(1)$). Then we obtain that

$$\left\| \hat{\mathbf{C}}_N - \hat{\mathbf{S}}'_N \right\| \xrightarrow{\text{a.s.}} 0$$

with

$$\hat{\mathbf{S}}'_N = v(\gamma) \frac{1}{n} \sum_{i=1}^{n-K} \mathbf{y}_i \mathbf{y}_i^\dagger + \frac{1}{n} \sum_{i=1}^{K} v(\alpha'_{i,n}) \mathbf{a}_i \mathbf{a}_i^\dagger$$

where $\alpha'_{i,n}$ are the unique positive solutions to

$$\alpha'_{i,n} = \frac{1}{N} \mathbf{a}_i^\dagger \left(\gamma^{-1} \mathbf{C}_N + \frac{1}{n} \sum_{j \neq i} v(\alpha'_{j,n}) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{a}_i.$$

As such, when the number K of outliers is fixed, the common weight $v(\gamma_n)$ becomes independent of the vectors $\mathbf{a}_i$'s (even if they are of arbitrarily large norm) while the individual weights $v(\alpha_{i,n})$ eventually solve a system of K equations involving the $\mathbf{a}_i$'s and $\mathbf{C}_N$.

A more specific case lies in the scenario where $\mathbf{a}_1 = \ldots = \mathbf{a}_K$. There, $\alpha_{1,n} = \ldots = \alpha_{K,n}$ and the K equations above reduce to a single one reading

$$\alpha'_{1,n} = \frac{1}{N} \mathbf{a}_1^\dagger \left(\gamma^{-1} \mathbf{C}_N + \frac{K-1}{n} v(\alpha'_{1,n}) \mathbf{a}_1 \mathbf{a}_1^\dagger \right)^{-1} \mathbf{a}_1$$

which, using $\mathbf{a}^\dagger (\mathbf{A} + t \mathbf{a} \mathbf{a}^\dagger)^{-1} = \mathbf{a}^\dagger \mathbf{A}^{-1} / (1 + t \mathbf{a}^\dagger \mathbf{A}^{-1} \mathbf{a})$ for invertible $\mathbf{A}$, simplifies as

$$\alpha'_{1,n} = \gamma \frac{\frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1}{1 + c_n \gamma (K-1) v(\alpha'_{1,n}) \frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1}$$

or equivalently

$$\frac{\alpha'_{1,n}}{1 - c_n (K-1) \psi(\alpha'_{1,n})} = \gamma \frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1.$$

Since the right-hand side is positive, so should be the left-hand side, which may then be seen as an increasing function of $\alpha'_{1,n}$. Thus, since γ depends neither on $\mathbf{C}_N$ nor $\mathbf{a}_1$, it comes that $\alpha'_{1,n}$ is an increasing function of $\frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1$. Moreover, $\alpha'_{1,n} < \psi^{-1}(1/(c_n(K-1)))$ and thus converges to zero as K grows large. When $K = 1$, and thus the outlier is now isolated, this reduces to

$$\alpha'_{1,n} = \gamma \frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1.$$

This short calculus leads to two important remarks. First, for $K = 1$, $\hat{\mathbf{C}}_N$ asymptotically allocates a weight $v(\gamma)$ to the legitimate data and a weight $v(\gamma \frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1)$ for the outlier. As a consequence, by the non-increasing property of v, the effect of the outlier will be (for most choices of the v function) attenuated if $\frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1 \ge 1$ but will be increased if $\frac{1}{N} \mathbf{a}_1^\dagger \mathbf{C}_N^{-1} \mathbf{a}_1 \le 1$. As such, the robust estimator of scatter will tend to mitigate the effect of outliers $\mathbf{a}_1$ having either large norm or, more interestingly, having strong alignment to the weakest eigenmodes of $\mathbf{C}_N$. In particular, note that when

$\mathbf{C}_N = \mathbf{I}_N$, $\hat{\mathbf{C}}_N$ will mostly control outliers upon their norms $\frac{1}{N}\|\mathbf{a}_1\|^2$, which is essentially what the normalized SCM

$$\frac{1}{n}\mathbf{Y}^{\mathrm{n}}\mathbf{Y}^{\mathrm{n}\dagger} = \frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} \frac{\mathbf{y}_i\mathbf{y}_i^\dagger}{\frac{1}{N}\|\mathbf{y}_i\|^2} + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} \frac{\mathbf{a}_i\mathbf{a}_i^\dagger}{\frac{1}{N}\|\mathbf{a}_i\|^2} \quad (5)$$

would do, and thus there is no gain in using robust estimators here. However, if $\mathbf{C}_N$ has large dimensional weak eigenspaces (i.e., close to singular with most eigenvalues near zero), $\frac{1}{N}\mathbf{a}_1^\dagger\mathbf{C}_N^{-1}\mathbf{a}_1$ may be quite large, and thus $\mathbf{a}_1$ may be strongly attenuated. But if $\mathbf{a}_1$ aligns to the strong eigenmodes of $\mathbf{C}_N$, the impact of $\mathbf{a}_1$ may be enhanced rather than reduced. To avoid this effect, undesirable in most cases, it is crucial to appropriately choose the u function. Specifically, the function v should be taken constant for all $x \leq \gamma$, or equivalently, $u(x)$ should be taken constant for $x \leq \phi^{-1}(1)$. A natural choice is the Huber estimator $u = u_{\mathrm{H}}$ introduced in (3).

The second remark is a slightly more surprising outcome. Indeed, despite n being potentially extremely large, the presence of (already few) $K > 1$ identical outliers drives $\hat{\mathbf{S}}_N$ (and thus $\hat{\mathbf{C}}_N$) to allocate large weights $v(\alpha_{i,n})$ (since $\alpha_{i,n}$ is small) to these outliers, therefore seemingly contradicting the very purpose of the robust estimator. This seems to indicate that $\hat{\mathbf{C}}_N$ has the propensity to put forward both large quantities of data with similar distribution *as well as* rather small quantities of vectors with strong pairwise alignment, while more naturally rejecting isolated outliers.

In terms of large dimensional spectral distribution and moments, the scenario of finitely many outliers is asymptotically equivalent to the outlier-free scenario. This can be observed from a rank-one perturbation argument along with $\varepsilon_n \to 0$ applied to Corollaries 1–2. A similar reasoning would hold for the normalized SCM. However, the matrices $\hat{\mathbf{C}}_N$ and $\frac{1}{n}\mathbf{Y}^{\mathrm{n}}\mathbf{Y}^{\mathrm{n}\dagger}$ themselves experience a (maximum) rank-K perturbation which can severely compromise the estimation of $\mathbf{C}_N$, along the previous argumentation lines.

Figure 1 displays an artificially generated scenario where a single outlier $\mathbf{a}_1$ of norm $\frac{1}{N}\|\mathbf{a}_1\|^2 = 1$ produces a large value for $\frac{1}{N}\mathbf{a}_1^\dagger\mathbf{C}_N^{-1}\mathbf{a}_1$ ($= 14.50$), thus entailing a strong attenuation by $\hat{\mathbf{C}}_N$. The terms $\mathbf{a}_1$ and $\mathbf{C}_N$ were made such that the SCM and normalized SCM have the same asymptotic eigenvalues and produce an isolated eigenvalue (around .25). The spectra of the latter are compared against those of $\hat{\mathbf{C}}_N$ and the oracle estimator. It is seen that the isolated eigenvalue, which is naturally not present in the spectrum of the oracle estimator, is also not present in the spectrum of $\hat{\mathbf{C}}_N$, indicating that $\hat{\mathbf{C}}_N$ has significantly reduced its impact on the spectrum.

Another interesting case study that shall provide further insight on $\hat{\mathbf{C}}_N$ is that where the $\mathbf{a}_i$'s (possibly numerous) are independently extracted from a different distribution to that of the $\mathbf{y}_i$'s. This is pursued in the subsequent section.

V. Random Outliers Scenario

Assuming $\mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n}$ to be independent with zero mean and covariance $\mathbf{D}_N \neq \mathbf{C}_N$ provides a rather immediate corollary of Theorem 1, given below. In the results to come, to differentiate between the conditions of Theorem 1 and those of Corollary 3, we shall use the subscript "R" standing for "random outliers scenario".

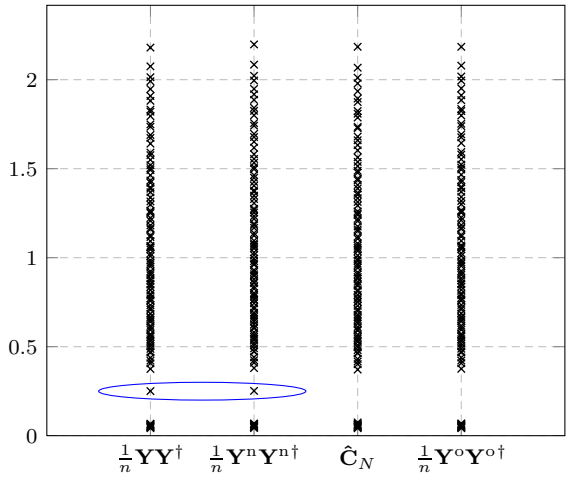


Fig. 1. Eigenvalues of the SCM ($\frac{1}{n}\mathbf{Y}\mathbf{Y}^\dagger$), normalized SCM ($\frac{1}{n}\mathbf{Y}^{\mathrm{n}}\mathbf{Y}^{\mathrm{n}\dagger}$), $\hat{\mathbf{C}}_N$ for $u = u_{\mathrm{S}}$ with $t = .1$, and the oracle estimator ($\frac{1}{n}\mathbf{Y}^{\mathrm{o}}\mathbf{Y}^{\mathrm{o}\dagger}$); $N = 100$, $c = .2$, $\varepsilon_n n = 1$, $\mathbf{a}_1 = (\mathbf{a}_1^1, \mathbf{a}_1^2)^\dagger$, $\mathbf{a}_1^1 \in \mathbb{R}^{10}$, $\mathbf{a}_1^2 \in \mathbb{R}^{90}$, with $\mathbf{a}_{1,i}^1 = \sqrt{10}$, $\mathbf{a}_{1,i}^2 = 0$, such that $\|\mathbf{a}_1\|^2 = N$; $\mathbf{y}_i = \mathbf{C}_N^{1/2}\mathbf{x}_i$ with $\mathbf{x}_{i,j}$ standard Gaussian and $\mathbf{C}_N = (16/14.50)\,\mathrm{diag}(\mathbf{c}_1, \mathbf{c}_2)$, $\mathbf{c}_1 \in \mathbb{R}^{10}$, $\mathbf{c}_2 \in \mathbb{R}^{90}$, with $\mathbf{c}_{1i} = 1/16$, $\mathbf{c}_{2i} = 1$, such that $\mathrm{tr}\,\mathbf{C}_N = N$. Ellipse around the outlier artifact.

Corollary 3 (Random Outliers)**.** *Let Assumption 2 hold with $\varepsilon > 0$ and let $\mathbf{a}_1, \ldots, \mathbf{a}_{\varepsilon_n n}$ be random independent of the $\mathbf{y}_i$'s with $\mathbf{a}_i = \mathbf{D}_N^{1/2}\mathbf{x}_i'$, where $\mathbf{D}_N \in \mathbb{C}^{N\times N}$ is deterministic Hermitian positive definite and $\mathbf{x}_1', \ldots, \mathbf{x}_{\varepsilon_n n}'$ are independent random vectors with i.i.d. zero mean, unit variance, and finite $(8+\eta)$-th order moment entries, for some $\eta > 0$. Let us further assume that $\limsup_N \|\mathbf{D}_N\mathbf{C}_N^{-1}\| < \infty$. Then, as $n \to \infty$,*

$$\left\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N^{\mathrm{R}}\right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{\mathbf{S}}_N^{\mathrm{R}} \triangleq v\left(\gamma_n^{\mathrm{R}}\right)\frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n}\mathbf{y}_i\mathbf{y}_i^\dagger + v\left(\alpha_n^{\mathrm{R}}\right)\frac{1}{n}\sum_{i=1}^{\varepsilon_n n}\mathbf{a}_i\mathbf{a}_i^\dagger$$

with γ_n^{R} and α_n^{R} the unique positive solutions to

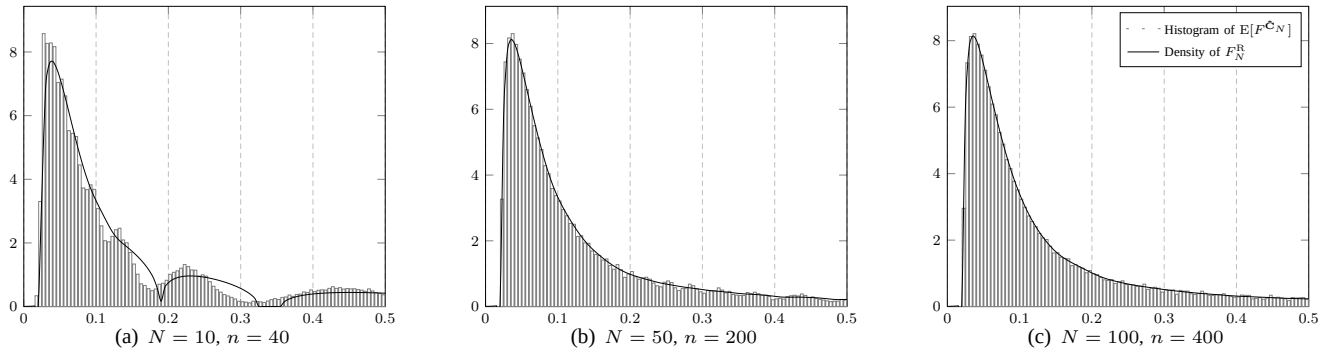
$$\gamma_n^{\mathrm{R}} = \frac{1}{N}\mathrm{tr}\,\mathbf{C}_N\left(\frac{(1-\varepsilon)v(\gamma_n^{\mathrm{R}})\mathbf{C}_N}{1+cv(\gamma_n^{\mathrm{R}})\gamma_n^{\mathrm{R}}} + \frac{\varepsilon v(\alpha_n^{\mathrm{R}})\mathbf{D}_N}{1+cv(\alpha_n^{\mathrm{R}})\alpha_n^{\mathrm{R}}}\right)^{-1}$$
$$\alpha_n^{\mathrm{R}} = \frac{1}{N}\mathrm{tr}\,\mathbf{D}_N\left(\frac{(1-\varepsilon)v(\gamma_n^{\mathrm{R}})\mathbf{C}_N}{1+cv(\gamma_n^{\mathrm{R}})\gamma_n^{\mathrm{R}}} + \frac{\varepsilon v(\alpha_n^{\mathrm{R}})\mathbf{D}_N}{1+cv(\alpha_n^{\mathrm{R}})\alpha_n^{\mathrm{R}}}\right)^{-1}.$$

In particular, for $F_N^{\hat{\mathbf{C}}_N}(x)$ as defined in Corollary 1,

$$F_N^{\hat{\mathbf{C}}_N}(x) - F_N^{\mathrm{R}}(x) \Rightarrow 0$$

almost surely as $n \to \infty$, where $F_N^{\mathrm{R}}(x)$ is a real distribution function with density, defined via its Stieltjes transform

$$m_N^{\mathrm{R}}(z) = \frac{1}{N}\mathrm{tr}\,\mathbf{E}_N^{-1}$$
$$\mathbf{E}_N = \frac{(1-\varepsilon)v(\gamma_n^{\mathrm{R}})}{1+e_{N,1}(z)}\mathbf{C}_N + \frac{\varepsilon v(\alpha_n^{\mathrm{R}})}{1+e_{N,2}(z)}\mathbf{D}_N$$



(a) $N = 10$, $n = 40$ (b) $N = 50$, $n = 200$ (c) $N = 100$, $n = 400$

Fig. 2. Density of F_N^{R} versus histogram of $\mathrm{E}[F^{\hat{\mathbf{C}}_N}]$ for $\mathbf{C}_N$ with $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, $\mathbf{D}_N = \mathbf{I}_N$, $\varepsilon = .05$, and $u(x) = (1+t)/(t+x)$ where $t = .1$.

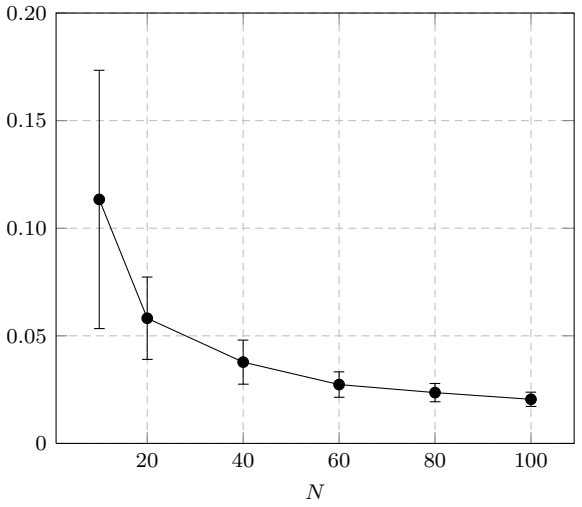


Fig. 3. Mean and standard deviation (error bars) of $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N^{\mathrm{R}}\|/\|\hat{\mathbf{C}}_N\|$ for $c_n = 0.25$, $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$, $\varepsilon_n = .05$, and $u(x) = (1+t)/(t+x)$, with $t = .1$.

for $z \in \mathbb{C}^+$ and $(e_{N,1}(z), e_{N,2}(z))$ the unique solution in $(\mathbb{C}^+)^2$ to

$$e_{N,1}(z) = \frac{v(\gamma_n^{\mathrm{R}})}{n} \operatorname{tr} \mathbf{C}_N \left(\mathbf{E}_N - z\mathbf{I}_N\right)^{-1}$$
$$e_{N,2}(z) = \frac{v(\alpha_n^{\mathrm{R}})}{n} \operatorname{tr} \mathbf{D}_N \left(\mathbf{E}_N - z\mathbf{I}_N\right)^{-1}.$$

Figure 2 shows the density of the distribution $\mathrm{E}[F_N^{\hat{\mathbf{C}}_N}]$, obtained from Monte-Carlo averaging, versus F_N^{R} for different values of N, n. It is observed that, as soon as N is of the order of several tens, the asymptotic approximation holds tightly. The (normalized) distance in spectral norm between $\hat{\mathbf{C}}_N$ and $\hat{\mathbf{S}}_N^{\mathrm{R}}$ is numerically evaluated in Figure 3 for various values of N. As suggested in the second order analysis of [11], $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\|$ (or $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N^{\mathrm{R}}\|$ here) is likely to decay at the rate $1/\sqrt{N}$, which is somewhat confirmed by observing that between $N = 20$ and $N = 80$, the approximation error decays by a factor of two (precisely, 0.042 versus 0.019).

In the random outliers scenario, $\hat{\mathbf{C}}_N$ is asymptotically equivalent to the weighted sum of two partial sample covariance matrices, one corresponding to the legitimate data and the other to the outlying data. In the defining equations for γ_n^{R} and α_n^{R} an interesting symmetrical interplay arises between the weights applied to the legitimate and the outlying data, which are only differentiated by ε. In particular, if $\varepsilon > 1/2$, the $\mathbf{a}_i$'s will be considered legitimate (being in majority) and the $\mathbf{y}_i$'s become outliers.

Despite the symmetrical form of the equations defining γ_n^{R} and α_n^{R}, it remains difficult to extract general insight on these quantities. Thus, again, it is interesting to study the regime where $\varepsilon \to 0$. In this case, $\gamma_n^{\mathrm{R}} \to \gamma = \phi^{-1}(1)/(1-c)$, and

$$\alpha_n^{\mathrm{R}} \to \gamma \frac{1}{N} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1}.$$

As such, the factor dictating the outlier mitigation strength of $\hat{\mathbf{C}}_N$ is now $\frac{1}{N} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1}$. Similar to before, when larger than one, the impact of the outliers will be reduced but these might be enhanced when smaller than one. Interestingly, if $\frac{1}{N} \operatorname{tr} \mathbf{D}_N = \frac{1}{N} \operatorname{tr} \mathbf{C}_N = 1$ (say), both legitimate and outlier samples have similar norm for all large n. As such, under this scenario, the SCM $\frac{1}{n}\mathbf{Y}\mathbf{Y}^\dagger$ or its normalized version $\frac{1}{n}\mathbf{Y}^{\mathrm{n}}\mathbf{Y}^{\mathrm{n}\dagger}$ behave asymptotically equivalently, neither of which being capable of differentiating between legitimate and outlier data. On the contrary, $\hat{\mathbf{C}}_N$ is capable of reducing the impact of the outliers as long as $\frac{1}{N} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1} > 1$. Note here again that $\mathbf{C}_N$ must be sufficiently distinct from $\mathbf{I}_N$, which would otherwise entail $\frac{1}{N} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1} \simeq 1$ and thus $\hat{\mathbf{C}}_N$ would be indifferent to outliers. Also, similar to previously, u must be well chosen to avoid enhancing the outlier effect if $\frac{1}{N} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1} < 1$ (so in particular it is advised that u be similar to u_{H}).

Figure 4 depicts the previous observations in terms of the deterministic equivalent spectral distributions: F_N^{R} of $\hat{\mathbf{C}}_N$, F_N^{SCM} of $\frac{1}{n}\mathbf{Y}\mathbf{Y}^\dagger$ (or F_N^{nSCM} of $\frac{1}{n}\mathbf{Y}^{\mathrm{n}}\mathbf{Y}^{\mathrm{n}\dagger}$ which satisfies $F_N^{\mathrm{SCM}} = F_N^{\mathrm{nSCM}}$ here), and F_N^{oracle} of the outlier-free oracle estimator $\frac{1}{n}\mathbf{Y}^{\mathrm{o}}\mathbf{Y}^{\mathrm{o}\dagger}$; we take here $\mathbf{C}_N$ and $\mathbf{D}_N$ to ensure $\frac{1}{n} \operatorname{tr} \mathbf{D}_N \mathbf{C}_N^{-1}$ large and ε is taken small. The sought-for distribution that would optimally discard all outliers is the oracle distribution and, thus, highly robust estimators are expected to have a similar distribution. Figure 4 confirms that this is indeed the case of $\hat{\mathbf{C}}_N$ which shows a close tail behavior but is slightly mismatched in the main distribution lobe. On the contrary, the SCM (normalized or not) shows a strong decay in the main lobe and a non matching tail. The associated theoretical values of γ_n^{R} and α_n^{R} for $\varepsilon_n = .05$ are here

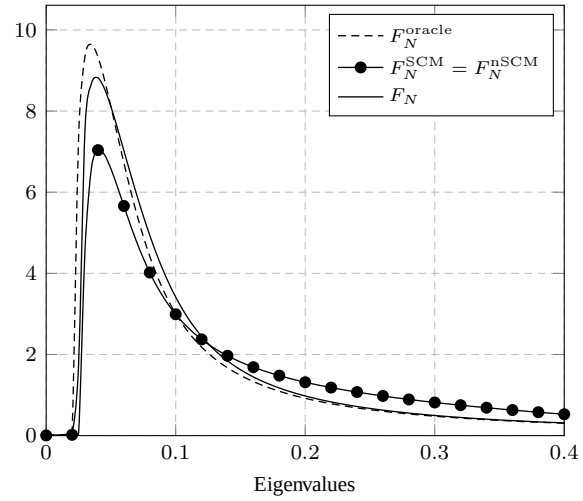


Fig. 4. Density of the approximate (deterministic) spectral distributions for the outlier-free oracle (F_N^{oracle}), the SCM or normalized SCM ($F_N^{\text{SCM}} = F_N^{\text{nSCM}}$), and $\hat{\mathbf{C}}_N$ (F_N), with $u = u_{\text{H}}$ with parameter $t = .1$, $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, $\mathbf{D}_N = \mathbf{I}_N$, $N = 100$, $c = .2$, and $\varepsilon = .05$.

$v_{\text{H}}(\gamma_n^{\text{R}}) \simeq 1.00$, $v_{\text{H}}(\alpha_n^{\text{R}}) \simeq .1219$, while in the limit $\varepsilon_n \to 0$, these values become $v_{\text{H}}(\gamma_n^{\text{R}}) \to 1$ and $v_{\text{H}}(\alpha_n^{\text{R}}) \to .1179$.

As it appears from Figure 4 that the tail of the various estimator distributions may be strongly affected by a weak outlier control, it is interesting to investigate the impact on their moments. For this, we introduce the following application to Corollary 2 for the random outlier setting.

Corollary 4 (Moments in Random Case)**.** *Under the setting of Corollary 1, letting $M_{N,p}^{\text{R}} = \int t^p dF_N^{\text{R}}(t)$, we have*

$$M_{N,p}^{\text{R}} = \frac{(-1)^p}{p!} \frac{1}{N} \operatorname{tr} \mathbf{T}_p^{\text{R}}$$

where $\mathbf{T}_p^{\text{R}}$ is obtained recursively as

$$\begin{aligned}
\mathbf{T}_{p+1}^{\text{R}} &= \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} \mathbf{T}_{p-i}^{\text{R}} \mathbf{Q}_{i-j+1}^{\text{R}} \mathbf{T}_j^{\text{R}} \\
\mathbf{Q}_{p+1}^{\text{R}} &= (p+1) \left[(1-\varepsilon) f_{1,p} \mathbf{R}_1 + \varepsilon f_{2,p} \mathbf{R}_2 \right] \\
f_{k,p+1} &= \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} (p-i+1) f_{k,j} f_{k,i-j} \beta_{k,p-i} \\
\beta_{k,p+1} &= \frac{1}{n} \operatorname{tr} \mathbf{R}_k \mathbf{T}_{p+1}^{\text{R}},
\end{aligned}$$

with initial values $\mathbf{T}_0^{\text{R}} = \mathbf{I}_N$, $f_{k,0} = -1$, $\beta_{k,0} = \frac{1}{n} \operatorname{tr} \mathbf{R}_k$, and with $\mathbf{R}_1 = v(\gamma_n^{\text{R}}) \mathbf{C}_N$, $\mathbf{R}_2 = v(\alpha_n^{\text{R}}) \mathbf{D}_N$. In particular,

$$\begin{aligned}
M_{N,1}^{\text{R}} &= \frac{1}{N} \operatorname{tr} \left[\varepsilon v(\alpha_n^{\text{R}}) \mathbf{D}_N + (1-\varepsilon) v(\gamma_n^{\text{R}}) \mathbf{C}_N \right] \\
M_{N,2}^{\text{R}} &= \frac{1}{N} \operatorname{tr} \Big[\left(\varepsilon v(\alpha_n^{\text{R}}) \mathbf{D}_N + (1-\varepsilon) v(\gamma_n^{\text{R}}) \mathbf{C}_N \right)^2 \\
&\quad + \varepsilon v^2(\alpha_n^{\text{R}}) \mathbf{D}_N \left[\frac{1}{n} \operatorname{tr} \mathbf{D}_N \right] + (1-\varepsilon) v^2(\gamma_n^{\text{R}}) \mathbf{C}_N \left[\frac{1}{n} \operatorname{tr} \mathbf{C}_N \right] \Big].
\end{aligned}$$

As expected, $\hat{\mathbf{C}}_N$ induces a bias in the mean. For fair comparison with the normalized SCM, which estimates $\mathbf{C}_N$ up to a scale constant, let us define the normalized moments

$$\bar{M}_{N,p} \triangleq \frac{M_{N,p}}{M_{N,1}}$$

and define similarly $\bar{M}_{N,p}^{\text{R}}$ as well as $\bar{M}_{N,p}^{\text{SCM}}$ for the SCM, $\bar{M}_{N,p}^{\text{nSCM}}$ for the normalized SCM, and $\bar{M}_{N,p}^{\text{oracle}}$ for the oracle estimator. Under the same setting as in Figure 4, we provide in the table of Figure 5 the successive normalized moments and relative error compared to $\bar{M}_{N,p}^{\text{oracle}}$. In this case, $\bar{M}_{N,p}^{\text{SCM}} = \bar{M}_{N,p}^{\text{nSCM}}$. For the scenario at end, given the large support of F_N^{R}, even low order moments tend to take large values so that the asymptotic moment approximation only theoretically holds for p rather small when $N = 100$ and we thus only provide these first order moments. The results demonstrate an important advantage brought by $\hat{\mathbf{C}}_N$ versus the SCM in that the first few order moments are better preserved.

	$p=2$	$p=3$	$p=4$
$\bar{M}_{N,p}^{\text{oracle}}$	9.28	129	1993
$\bar{M}_{N,p}^{\text{R}}$ (error)	9.18 (1.1%)	126 (1.8%)	1945 (2.4%)
$\bar{M}_{N,p}^{\text{SCM}}$ (error)	8.53 (8.2%)	112 (13%)	1660 (17%)

Fig. 5. Normalized moments $\bar{M}_{N,p}^{\text{R}}$, $\bar{M}_{N,p}^{\text{SCM}}$, versus $\bar{M}_{N,p}^{\text{oracle}}$, and relative error $|\cdot - \bar{M}_{N,p}^{\text{oracle}}|/\bar{M}_{N,p}^{\text{oracle}}$. Random outliers, $N = 100$, $c = .2$, $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, $\mathbf{D}_N = \mathbf{I}_N$, $\varepsilon = .05$, $u = u_{\text{H}}$, $t = .1$.

VI. Discussion and Concluding Remarks

Our study of the robust estimator $\hat{\mathbf{C}}_N$ in the large random matrix regime has already led to several interesting conclusions, which we shall more thoroughly address in this section.

Most investigations of robust estimators of scatter focus on the more tractable case where the samples (i.e., the columns of $\mathbf{Y}$) are independent with identical elliptical distribution. The recent results of [9, 10] have revealed that, as $u(x)$ gets close to the Tyler $1/x$ function, in the large random matrix regime, $\hat{\mathbf{C}}_N$ tends to behave similar to the normalized SCM defined in (5). This conclusion was quite pessimistic as it suggested no real improvement of $\hat{\mathbf{C}}_N$ over simplistic alternative robust methods. In the concluding remarks of [10, Section 4], the authors anticipated a change of behavior of $\hat{\mathbf{C}}_N$ versus the normalized SCM for deterministic outlier data. This was revealed here both in Section IV and in Section V where it is made clear that, unlike the normalized SCM, the robust estimators of scatter smartly detect the outliers, essentially by evaluating and comparing the quadratic forms $\mathbf{y}^\dagger \mathbf{C}_N^{-1} \mathbf{y}$ for each column vector $\mathbf{y}$ of $\mathbf{Y}$. Larger $\mathbf{y}^\dagger \mathbf{C}_N^{-1} \mathbf{y}$ imply more attenuation of $\mathbf{y}$ within the observed samples. However, an incidental consequence of this behavior of $\hat{\mathbf{C}}_N$ is that small values of $\mathbf{y}^\dagger \mathbf{C}_N^{-1} \mathbf{y}$ enhance the effect of $\mathbf{y}$ even though it might not comply with the legitimate sample distribution, thus increasing the probability of inducing false alarms. This has led us to conclude that the function u should be adequately tuned to avoid such a phenomenon. Another consequence is that matrices $\hat{\mathbf{C}}_N$ with legitimate data of covariance $\mathbf{C}_N$ close to the identity will have very poor outlier rejection properties.

When the outliers are few, the empirical spectral measure $F^{\hat{\mathbf{C}}_N}$ of $\hat{\mathbf{C}}_N$ is asymptotically the same as that of the SCM,

normalized SCM, and oracle estimators. As such, if one's interest is on functionals of the eigenvalues of $\mathbf{C}_N$, such as moments, and only few outliers are expected, sophisticated robust estimators come to no avail. This being said, the outliers may naturally engender extra isolated eigenvalues (only finitely many) in the spectrum of $\frac{1}{n}\mathbf{Y}\mathbf{Y}^\dagger$ which $\hat{\mathbf{C}}_N$ might suitably remove while the normalized SCM may not (recall Figure 1). For subspace detection and estimation applications, where the information often lies in the eigenvectors of isolated eigenvalues, discarding such outlying information is critical and thus robust estimators may bring important performance gains. For instance, applications in finance and biostatistics (where data are often assumed to contain outliers) heavily rely on isolated eigenvalue-eigenvector pairs, see e.g., [20, 21]. The experimenter must however keep in mind that, according to our analysis, $\hat{\mathbf{C}}_N$ is most effective at automatically suppressing *isolated* outliers (the less of these relative to the legitimate samples the better) and loses discriminatory power as the outliers approach one another.

The observation made in Section V that the distribution (in particular through its first order moments) F_N^{R} is much closer to the oracle estimator than would the (normalized or not) SCM be leads to some interesting applications when it comes to designing improved estimators for $\mathbf{C}_N$ that both account for the fact that n is not large compared to N and for the fact that the observed data are prone to outliers. Such investigations were successively made in [22] for the finite N, n regime and later in [10] for the large N, n regime where hybrid Ledoit–Wolf [23] and Tyler [8] estimators were proposed that improve the estimation of $\mathbf{C}_N$ by providing an extra degree of freedom (a regularization parameter) which is selected so to minimize the expected Frobenius norm between $\mathbf{C}_N$ and the estimator under study. Since the Frobenius norm is nothing but a functional of second order moments, the observation made in the table of Figure 5 strongly suggests that the Ledoit–Wolf estimator alone (being based on the SCM) would be quite sensitive to deterministic outliers while the estimators studied in [10, 22], which are essentially of a similar class as $\hat{\mathbf{C}}_N$, would be much more resilient to such outliers.

When the number of outliers is much larger, even in the random outlier scenario studied in Section V, very little can be said. However, we noticed an interesting symmetry in the equations defining the weights γ_n^{R} and α_n^{R} of Corollary 3, which reveals that the asymptotic proportion ε of outliers versus $1-\varepsilon$ of legitimate data could tip for $\varepsilon > .5$ towards letting the outliers be considered as the truly legitimate data.

In summary, the present study provides a first step towards a better understanding of the behavior of (classical) robust estimators of scatter against arbitrary outliers. Our findings underline several key aspects of such estimators of profound practical relevance, such as the importance of the population covariance matrix $\mathbf{C}_N$ of the legitimate data in the rejection power of the estimator, as well as the risks inherent to using weight functions u of the Tyler type. Nonetheless, this study remains at the theoretical level of the estimator itself and does not consider the implications when used as a plug-in estimator in detection or estimation methods. Whether these methods are based on local information (isolated eigenvalue, specific eigenvectors, etc.) or global information (functional of the eigenvalues, projections on large subspaces, etc.) about $\mathbf{C}_N$ will entail significant differences in the way $\hat{\mathbf{C}}_N$, through the weight function u, must be tailored. Such considerations are left to future investigations.

Appendix A
Proof of Theorem 1

The main technical difficulty of the article lies in the proof of Theorem 1 which extends the methods developed in [9] to multiple sample types. The present section is dedicated to this proof. Some auxiliary random matrix results will be then listed in Appendix B, while Appendix C will deal with the (rather immediate) proof of Corollary 2.

The proof of Theorem 1 is divided in two parts. First, we show that the system of fixed-point equations (4) admits a unique vector solution and that such solution is bounded as $n \to \infty$. This then defines unequivocally the matrix $\hat{\mathbf{S}}_N$. We then show in a second part that $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\| \xrightarrow{\text{a.s.}} 0$.

A. Existence, uniqueness, boundedness of the solution to (4)

To prove existence and uniqueness, we use the framework of standard interference functions [24].

Definition 1. *A function $\mathbf{h} = (h_0, \ldots, h_s) : \mathbb{R}_+^{1+s} \to \mathbb{R}_+^{1+s}$ is a standard interference function if it satisfies the conditions:*

1) **Positivity***: if $q_0, \ldots, q_s \geq 0$, then $h_i(q_0, \ldots, q_s) > 0$ for all i.*
2) **Monotonicity***: if $q_0 \geq q_0', \ldots, q_s \geq q_s'$ then, for all i, $h_i(q_0, \ldots, q_s) \geq h_i(q_0', \ldots, q_s')$.*
3) **Scalability***: for all $\delta > 1$ and all i, $\delta h_i(q_0, \ldots, q_s) > h_i(\delta q_0, \ldots, \delta q_s)$.*

By [24, Thm. 2], if $\mathbf{h}$ is a standard interference function for which there exists $(q_0, \ldots, q_s)$ such that $q_i \geq h_i(q_0, \ldots, q_s)$ for all i, then the system of equations $q_i = h_i(q_0, \ldots, q_s)$, $i = 0, \ldots, s$, has a unique solution.

Define $\mathbf{h} \triangleq (h_0, \ldots, h_{\varepsilon_n n}) : \mathbb{R}_+^{1+\varepsilon_n n} \to \mathbb{R}_+^{1+\varepsilon_n n}$ with

$$h_0(q_0, \ldots, q_{\varepsilon_n n}) = \frac{1}{N} \operatorname{tr} \mathbf{C}_N \left(\frac{(1-\varepsilon)v(q_0)}{1 + cv(q_0)q_0} \mathbf{C}_N + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v(q_j)\, \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1}$$

$$h_i(q_0, \ldots, q_{\varepsilon_n n}) = \frac{1}{N} \mathbf{a}_i^\dagger \left(\frac{(1-\varepsilon)v(q_0)}{1 + cv(q_0)q_0} \mathbf{C}_N + \frac{1}{n} \sum_{j \neq i} v(q_j)\, \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{a}_i$$

for $i = 1, \ldots, \varepsilon_n n$. Let us prove that $\mathbf{h}$ meets the conditions of Definition 1 and that, for $i = 0, \ldots, \varepsilon_n n$, $h_i(q_0, \ldots, q_{\varepsilon_n n}) \leq q_i$ for some $(q_0, \ldots, q_{\varepsilon_n n})$, which will then prove existence and uniqueness.

From Assumption 1 and the fact that v is bounded, we clearly have $h_i > 0$ for all i. To show monotonicity, let us

first define

$$\mathbf{B}_N(q_0,\ldots,q_{\varepsilon_n n}) = \frac{(1-\varepsilon)v(q_0)}{1+cv(q_0)q_0}\mathbf{C}_N + \frac{1}{n}\sum_{j=1}^{\varepsilon_n n} v\left(q_j\right)\mathbf{a}_j\mathbf{a}_j^\dagger$$

and take $q_0,\ldots,q_{\varepsilon_n n}$ and $q_0',\ldots,q_{\varepsilon_n n}'$ such that $q_i \geq q_i'$ for all i. Then, since v is non-increasing and $\psi(x) = xv(x)$ is increasing,

$$\mathbf{B}_N(q_0,\ldots,q_{\varepsilon_n n}) \preceq \mathbf{B}_N(q_0',\ldots,q_{\varepsilon_n n}').$$

From [17, Cor. 7.7.4], this implies

$$\left(\mathbf{B}_N(q_0,\ldots,q_{\varepsilon_n n})\right)^{-1} \succeq \left(\mathbf{B}_N(q_0',\ldots,q_{\varepsilon_n n}')\right)^{-1}$$

from which $h_0(q_0,\ldots,q_{\varepsilon_n n}) \geq h_0(q_0',\ldots,q_{\varepsilon_n n}')$. By the same arguments, $h_i(q_0,\ldots,q_{\varepsilon_n n}) \geq h_i(q_0',\ldots,q_{\varepsilon_n n}')$ for $i = 1,\ldots,\varepsilon_n n$, thus proving the monotonicity of $\mathbf{h}$. Finally, to show scalability, let us rewrite h_0 as

$$h_0(q_0,\ldots,q_{\varepsilon_n n}) = \frac{1}{N}\operatorname{tr}\mathbf{C}_N\left((1-\varepsilon)\frac{\Theta(q_0)}{q_0}\mathbf{C}_N + \frac{1}{n}\sum_{j=1}^{\varepsilon_n n}\frac{\psi\left(q_j\right)}{q_j}\mathbf{a}_j\mathbf{a}_j^\dagger\right)^{-1}$$

where $\Theta(x) = \frac{\psi(x)}{1+c\psi(x)}$. Since $\psi(x)$ is increasing, so is $\Theta(x)$ and, for any $\delta > 1$,

$$\begin{aligned} &h_0(\delta q_0,\ldots,\delta q_{\varepsilon_n n}) \\ &= \frac{\delta}{N}\operatorname{tr}\mathbf{C}_N\left(\frac{(1-\varepsilon)\Theta(\delta q_0)}{q_0}\mathbf{C}_N + \frac{1}{n}\sum_{j=1}^{\varepsilon_n n}\frac{\psi\left(\delta q_j\right)}{q_j}\mathbf{a}_j\mathbf{a}_j^\dagger\right)^{-1} \\ &< \delta h_0(q_0,\ldots,q_{\varepsilon_n n}). \end{aligned}$$

We show similarly $h_i(\delta q_0,\ldots,\delta q_{\varepsilon_n n}) < \delta h_i(q_0,\ldots,q_{\varepsilon_n n})$ for $i = 1,\ldots,\varepsilon_n n$, thus proving the scalability of $\mathbf{h}$.

Thus, $\mathbf{h}$ is a standard interference function and it remains to show that $h_i(q_0,\ldots,q_{\varepsilon_n n}) \leq q_i$ for some $(q_0,\ldots,q_{\varepsilon_n n})$ and for all i. For $i = 0$,

$$h_0(q_0,\ldots,q_{\varepsilon_n n}) = \frac{1}{N}\operatorname{tr}\mathbf{C}_N\left(\mathbf{B}_N(q_0,\ldots,q_{\varepsilon_n n})\right)^{-1}$$

where

$$\mathbf{B}_N(q_0,\ldots,q_{\varepsilon_n n}) \succeq \frac{(1-\varepsilon)v(q_0)}{1+cv(q_0)q_0}\mathbf{C}_N$$

and thus, by definition of ψ,

$$h_0(q_0,\ldots,q_{\varepsilon_n n}) \leq \frac{1+c\psi(q_0)}{(1-\varepsilon)\psi(q_0)}q_0. \tag{6}$$

As a consequence, we need to find some q_0 for which $\frac{1+c\psi(q_0)}{(1-\varepsilon)\psi(q_0)} \leq 1$ or, equivalently, $\psi(q_0) \geq \frac{1}{1-\varepsilon-c}$. Such a choice of q_0 is always possible since ψ is increasing on $[0,\infty)$ with image $[0,\psi_\infty)$ where $\frac{1}{1-\varepsilon-c} < \psi_\infty$ (this unfolds from $\phi_\infty > \frac{1}{1-\varepsilon}$). Therefore, for any q_0 such that $\frac{1}{1-\varepsilon-c} \leq \psi(q_0) < \psi_\infty$, we have $h_0(q_0,\ldots,q_{\varepsilon_n n}) \leq q_0$. Take for instance $q_0 = \psi^{-1}(\frac{1}{1-\varepsilon-c})$ and consider now the functions h_i, $i = 1,\ldots,\varepsilon_n n$ for which, using [25, Lemma 10] and similar arguments as above,

$$\begin{aligned} h_i(q_0,\ldots,q_{\varepsilon_n n}) &\leq q_0\frac{1+c\psi(q_0)}{(1-\varepsilon)\psi(q_0)}\frac{1}{N}\mathbf{a}_i^\dagger\mathbf{C}_N^{-1}\mathbf{a}_i \\ &= q_0\frac{1}{N}\mathbf{a}_i^\dagger\mathbf{C}_N^{-1}\mathbf{a}_i \triangleq w_i. \end{aligned} \tag{7}$$

Therefore, taking $q_i = w_i$ for $i = 1,\ldots,\varepsilon_n n$, we also have $h_i(q_0,\ldots,q_{\varepsilon_n n}) \leq q_i$. Altogether, we have shown that the function $\mathbf{h}$ satisfies the conditions of [24, Thm. 2] implying that there exists a unique solution to (4). As such, $\hat{\mathbf{S}}_N$ as introduced in the statement of Theorem 1 is well-defined.

We now turn our focus to the boundedness of the solution to (4). From (6) and (7), along with Assumption 1, we immediately have that $(\gamma_n, \alpha_{1,n},\ldots,\alpha_{\varepsilon_n n,n})$ is uniformly bounded in n, i.e., $\limsup_n \gamma_n < \infty$ and $\limsup_n \max_{1\leq i\leq \varepsilon_n n}\alpha_{i,n} < \infty$. Furthermore, γ_n can be shown to be uniformly away from zero as follows. By monotonicity of the $\mathbf{h}$ function, $h_0(q_0,\ldots,q_{\varepsilon_n n}) \geq h_0(0,\ldots,0)$, i.e.,

$$h_0(q_0,\ldots,q_{\varepsilon_n n}) \geq \frac{1}{v(0)}\frac{1}{N}\operatorname{tr}\mathbf{H}_N^{-1} \geq \frac{1}{v(0)}\frac{1}{\|\mathbf{H}_N\|},$$

where the matrix $\mathbf{H}_N$ is defined as

$$\mathbf{H}_N \triangleq (1-\varepsilon)\mathbf{I}_N + \frac{1}{n}\sum_{j=1}^{\varepsilon_n n}\mathbf{C}_N^{-1/2}\mathbf{a}_j\mathbf{a}_j^\dagger\mathbf{C}_N^{-1/2}.$$

By Assumption 1 we have $\limsup_n \|\mathbf{H}_N\| < \infty$ and, consequently, $\liminf_n \gamma_n > 0$.

B. Convergence of $\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N$

Having proved that $\hat{\mathbf{S}}_N$ is well defined, we now turn to the core of the proof of Theorem 1. The outline of the proof follows tightly that of [9, Thm. 2] but for a model that is (i) simpler in its assuming the legitimate data to be essentially Gaussian instead of elliptical, but (ii) made more complex due to the deterministic addition of the vectors $\mathbf{a}_1,\ldots,\mathbf{a}_{\varepsilon_n n}$. Our way to deal with (ii) is by controlling in parallel the quantities asymptotically approximated by γ_n and those asymptotically approximated by $\alpha_{i,n}$. Since some parts of the proof mirror closely those in [9, Thm. 2], we shall mainly focus on the significantly differing aspects.

First note that we can assume $\mathbf{C}_N = \mathbf{I}_N$ by studying $\mathbf{C}_N^{-1/2}\hat{\mathbf{C}}_N\mathbf{C}_N^{-1/2}$ instead of $\hat{\mathbf{C}}_N$, in which case we have $\mathbf{C}_N^{-1/2}\mathbf{a}_i$ in place of the original $\mathbf{a}_i$. This can be seen from (1), the implicit equation solved by $\hat{\mathbf{C}}_N$. Hence, from now on we assume $\mathbf{C}_N = \mathbf{I}_N$ without loss of generality. Using the definition $v(x) \triangleq u\left(g_n^{-1}(x)\right)$, with $g_n(x) = x/(1-c_n\phi(x))$, and following the same steps as in [9], let us write

$$\hat{\mathbf{C}}_N = \frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} v\left(d_i\right)\mathbf{x}_i\mathbf{x}_i^\dagger + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} v\left(b_i\right)\mathbf{a}_i\mathbf{a}_i^\dagger$$

with $d_i \triangleq \frac{1}{N}\mathbf{x}_i^\dagger\hat{\mathbf{C}}_{(x_i)}^{-1}\mathbf{x}_i$ and $b_i \triangleq \frac{1}{N}\mathbf{a}_i^\dagger\hat{\mathbf{C}}_{(a_i)}^{-1}\mathbf{a}_i$, where $\hat{\mathbf{C}}_{(x_i)} \triangleq \hat{\mathbf{C}}_N - v\left(d_i\right)\mathbf{x}_i\mathbf{x}_i^\dagger$ and $\hat{\mathbf{C}}_{(a_i)} \triangleq \hat{\mathbf{C}}_N - v\left(b_i\right)\mathbf{a}_i\mathbf{a}_i^\dagger$. Further define

$$e_i \triangleq \frac{v(d_i)}{v(\gamma_n)}, \qquad f_i \triangleq \frac{v(b_i)}{v(\alpha_{i,n})},$$

with γ_n and $\alpha_{i,n}$ as in the statement of Theorem 1 but for

$\mathbf{C}_N = \mathbf{I}_N$, i.e., γ_n and $\alpha_{i,n}$ are the positive solutions to

$$\gamma_n = \frac{1}{N} \operatorname{tr} \left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} \mathbf{I}_N + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v\left(\alpha_{j,n}\right) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1}$$

$$\alpha_{i,n} = \frac{1}{N} \mathbf{a}_i^\dagger \left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} \mathbf{I}_N + \frac{1}{n} \sum_{j \neq i} v\left(\alpha_{j,n}\right) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{a}_i.$$

The core of the proof is to show that

$$\max_{1 \leq i \leq (1-\varepsilon_n)n} |e_i - 1| \xrightarrow{\text{a.s.}} 0 \tag{8}$$

$$\max_{1 \leq i \leq \varepsilon_n n} |f_i - 1| \xrightarrow{\text{a.s.}} 0. \tag{9}$$

Let us first relabel e_i and f_i such that $e_1 \leq \ldots \leq e_{(1-\varepsilon_n)n}$ and $f_1 \leq \ldots \leq f_{\varepsilon_n n}$ and denote $\delta_n = \max(e_{(1-\varepsilon_n)n}, f_{\varepsilon_n n})$. For any $i = 1, \ldots, (1-\varepsilon_n)n$, we have

$$e_i = \frac{v\left(\frac{1}{N} \mathbf{x}_i^\dagger \left(\frac{1}{n} \sum_{j \neq i} v(d_j) \mathbf{x}_j \mathbf{x}_j^\dagger + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v(b_j) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{x}_i \right)}{v(\gamma_n)}$$

$$\leq \frac{v\left(\frac{1}{\delta_n N} \mathbf{x}_i^\dagger \left(\frac{1}{n} \sum_{j \neq i} v(\gamma_n) \mathbf{x}_j \mathbf{x}_j^\dagger + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v(\alpha_{j,n}) \mathbf{a}_j \mathbf{a}_j^\dagger \right)^{-1} \mathbf{x}_i \right)}{v(\gamma_n)}$$

where we used $v(d_j) = v(\gamma_n) e_j$, $v(b_j) = v(\alpha_{j,n}) f_j$ and the inequality arises from $e_j, f_j \leq \delta_n$, from v being non-increasing, and from [17, Cor. 7.7.4]. For readability, let

$$\mathbf{F}_{N,(i)} \triangleq \frac{1}{n} \sum_{j \neq i} v(\gamma_n) \mathbf{x}_j \mathbf{x}_j^\dagger + \frac{1}{n} \sum_{j=1}^{\varepsilon_n n} v(\alpha_{j,n}) \mathbf{a}_j \mathbf{a}_j^\dagger.$$

From the random matrix result, Lemma 1 of Appendix B,

$$\max_{1 \leq i \leq (1-\varepsilon_n)n} \left| \frac{1}{N} \mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1} \mathbf{x}_i - \gamma_n \right| \xrightarrow{\text{a.s.}} 0.$$

Thus, for $\zeta > 0$, with probability one, we have for all large n

$$e_{(1-\varepsilon_n)n} \leq \frac{v\left(\frac{1}{\delta_n}(\gamma_n - \zeta) \right)}{v(\gamma_n)}. \tag{10}$$

We can proceed similarly to bound f_i from above as

$$f_i \leq \frac{v\left(\frac{1}{\delta_n N} \mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1} \mathbf{a}_i \right)}{v(\alpha_{i,n})}$$

for any $i = 1, \ldots, \varepsilon_n n$, with

$$\mathbf{G}_{N,(i)} \triangleq \frac{1}{n} \sum_{j=1}^{(1-\varepsilon_n)n} v\left(\gamma_n\right) \mathbf{x}_j \mathbf{x}_j^\dagger + \frac{1}{n} \sum_{j \neq i} v\left(\alpha_{j,n}\right) \mathbf{a}_j \mathbf{a}_j^\dagger$$

and we now use Lemma 2 in Appendix B which states

$$\max_{1 \leq i \leq \varepsilon_n n} \left| \frac{1}{N} \mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1} \mathbf{a}_i - \alpha_{i,n} \right| \xrightarrow{\text{a.s.}} 0.$$

Therefore, for the same $\zeta > 0$ and for all large n a.s.,

$$f_{\varepsilon_n n} \leq \frac{v\left(\frac{1}{\delta_n}(\alpha_{i,n} - \zeta) \right)}{v(\alpha_{i,n})}. \tag{11}$$

We now consider separately the subsequence of n over which $e_{(1-\varepsilon_n)n} \geq f_{\varepsilon_n n}$ and that over which $e_{(1-\varepsilon_n)n} < f_{\varepsilon_n n}$ (these subsequences may be empty or finite).

Subsequence $e_{(1-\varepsilon_n)n} \geq f_{\varepsilon_n n}$: On this subsequence, (10) becomes

$$e_{(1-\varepsilon_n)n} \leq \frac{v\left(\frac{1}{e_{(1-\varepsilon_n)n}}(\gamma_n - \zeta) \right)}{v(\gamma_n)}$$

or alternatively, since $e_{(1-\varepsilon_n)n}$ is positive,

$$1 \leq \frac{\psi\left(\frac{\gamma_n}{e_{(1-\epsilon)n}} \left(1 - \frac{\zeta}{\gamma_n}\right) \right)}{\psi(\gamma_n)\left(1 - \frac{\zeta}{\gamma_n}\right)}.$$

We want to prove that, for any $\ell > 0$, $e_{(1-\epsilon)n} \leq 1 + \ell$ for all large n a.s. Let us assume the opposite, i.e., $e_{(1-\epsilon)n} > 1+\ell$ infinitely often, and let us restrict ourselves to a (further) subsequence where this always holds. Then,

$$1 \leq \frac{\psi\left(\frac{\gamma_n}{1+\ell} \left(1 - \frac{\zeta}{\gamma_n}\right) \right)}{\psi(\gamma_n)\left(1 - \frac{\zeta}{\gamma_n}\right)} \leq \frac{\psi\left(\frac{\gamma_n}{1+\ell} \right)}{\psi(\gamma_n)\left(1 - \frac{\zeta}{\gamma_n}\right)}.$$

From the uniform boundedness of γ_n away from zero and infinity (see Appendix A-A), considering yet a further subsequence over which $\gamma_n \to \gamma_0 > 0$, we obtain in the limit

$$\psi(\gamma_0)\left(1 - \frac{\zeta}{\gamma_0}\right) \leq \psi\left(\frac{\gamma_0}{1+\ell} \right).$$

This being valid for each $\zeta > 0$, a contradiction is raised in the limit $\zeta \to 0$. Therefore, either the subsequence over which $e_{(1-\varepsilon_n)n} \geq f_{\varepsilon_n n}$ is finite or $e_{(1-\epsilon)n} \leq 1 + \ell$ for all large n a.s. Assuming the former, then $e_{(1-\varepsilon_n)n} < f_{\varepsilon_n n}$ for all large n, which is considered next.

Subsequence $e_{(1-\varepsilon_n)n} < f_{\varepsilon_n n}$: On this subsequence, (11) becomes

$$f_{\varepsilon_n n} \leq \frac{v\left(\frac{1}{f_{\varepsilon_n n}}(\alpha_{\varepsilon_n n, n} - \zeta) \right)}{v(\alpha_{\varepsilon_n n, n})} \tag{12}$$

for all large n a.s. Again, we wish to prove that with, say, the same $\ell > 0$ as above, $f_{\varepsilon_n n} \leq 1+\ell$ for all large n a.s. Consider first the case $\liminf_n \alpha_{\varepsilon_n n, n} = 0$ and restrict ourselves to those converging subsequences over which $\alpha_{\varepsilon_n n, n} \to 0$. In the limit, $v(\alpha_{\varepsilon_n n, n}) \to v(0)$ so that, for any $\theta > 0$ and for n large enough, $v(\alpha_{\varepsilon_n n, n}) > v(0) - \theta$. This, along with $v(1/f_{\varepsilon_n n}(\alpha_{\varepsilon_n n, n} - \zeta)) \leq v(0)$ gives $f_n \leq v(0)/(v(0) - \theta)$ for all large n implying that, for any $\ell > 0$, $f_n \leq 1 + \ell$ for all large n a.s. Consider now the rest of subsequences for which $\liminf_n \alpha_{\varepsilon_n n, n} > 0$ and rewrite (12) as

$$1 \leq \frac{\psi\left(\frac{\alpha_{\varepsilon_n n, n}}{f_{\varepsilon_n n}} \left(1 - \frac{\zeta}{\alpha_{\varepsilon_n n, n}}\right) \right)}{\psi(\alpha_{\varepsilon_n n, n})\left(1 - \frac{\zeta}{\alpha_{\varepsilon_n n, n}}\right)}.$$

As above for $e_{(1-\varepsilon_n)n}$, we assume $f_{\varepsilon_n n} > 1 + \ell$ infinitely often, and restrict ourselves to a further subsequence where this holds for all n. Then,

$$1 \leq \frac{\psi\left(\frac{\alpha_{\varepsilon_n n, n}}{1+\ell} \right)}{\psi(\alpha_{\varepsilon_n n, n})\left(1 - \frac{\zeta}{\alpha_{\varepsilon_n n, n}}\right)}.$$

From the boundedness of $\alpha_{\varepsilon_n n,n}$ (see Appendix A-A), we can take a converging (further) subsequence over which $\alpha_{\varepsilon_n n,n} \to \alpha_0 > 0$. In the limit,

$$\psi(\alpha_0)\left(1-\frac{\zeta}{\alpha_0}\right) \leq \psi\left(\frac{\alpha_0}{1+\ell}\right)$$

which is contradictory for sufficiently small ζ. Thus, necessarily $f_{\varepsilon_n n} \leq 1+\ell$ for all large n a.s., unless we have $e_{(1-\varepsilon_n)n)} \geq f_{\varepsilon_n n}$ in which case, as shown above, $f_{\varepsilon_n n} \leq e_{(1-\varepsilon_n)n)} \leq 1+\ell$ for all large n a.s.

Altogether, we necessarily have

$$\max\{e_{(1-\varepsilon_n)n}, f_{\varepsilon_n n}\} \leq 1+\ell$$

for all large n a.s. All the same, by reverting the inequalities, we prove that, for all large n a.s.

$$\min\{e_1, f_1\} \geq 1-\ell$$

and therefore, altogether,

$$\max_{1\leq i\leq (1-\epsilon)n} |e_i - 1| \leq \ell$$
$$\max_{1\leq i\leq \epsilon n} |f_i - 1| \leq \ell$$

for all large n a.s., which eventually proves (8) and (9) by taking a countable sequence of ℓ going to zero. This establishes the main result, from which Theorem 1 unfolds. Specifically, from (8)-(9) and by uniform boundedness of γ_n and $\alpha_{i,n}$,

$$\max_{1\leq i\leq (1-\varepsilon_n)n} |v(d_i) - v(\gamma_n)| \xrightarrow{\text{a.s.}} 0$$
$$\max_{1\leq i\leq \varepsilon_n n} |v(b_i) - v(\alpha_{i,n})| \xrightarrow{\text{a.s.}} 0.$$

Thus, for any $\ell > 0$ and for all large n a.s.

$$(1-\ell)\hat{\mathbf{S}}_N \preceq \hat{\mathbf{C}}_N \preceq (1+\ell)\hat{\mathbf{S}}_N$$

and, therefore, $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\| \leq 2\ell\|\hat{\mathbf{S}}_N\|$. Using the triangle inequality and the fact that v is non-increasing, we have

$$\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\| \leq 2\ell\, v(0)\left(\left\|\frac{1}{n}\sum\nolimits_{i=1}^{(1-\varepsilon_n)n} \mathbf{x}_i\mathbf{x}_i^\dagger\right\| + \left\|\frac{1}{n}\sum\nolimits_{i=1}^{\varepsilon_n n} \mathbf{a}_i\mathbf{a}_i^\dagger\right\|\right).$$

From [26] and Assumption 2, $\|\frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} \mathbf{x}_i\mathbf{x}_i^\dagger\| < 4(1-\varepsilon)$ for all large n a.s. and, from Assumption 1, $\limsup_n \|\frac{1}{n}\sum_{i=1}^{\varepsilon_n n} \mathbf{a}_i\mathbf{a}_i^\dagger\| < \infty$. Then, since ℓ is arbitrarily small, $\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\|$ tends to zero a.s. as $n \to \infty$, which concludes the proof of Theorem 1. For $\mathbf{C}_N \neq \mathbf{I}_N$ we simply need to show $\|\mathbf{C}_N^{1/2}(\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N)\mathbf{C}_N^{1/2}\| \xrightarrow{\text{a.s.}} 0$, which follows from $\|\mathbf{C}_N^{1/2}(\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N)\mathbf{C}_N^{1/2}\| \leq \|\mathbf{C}_N\|\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\|$ since, by assumption, $\limsup_N \|\mathbf{C}_N\| < \infty$.

For the random outliers scenario, Assumption 1 holds a.s. by virtue of [26], provided that $\limsup_N \|\mathbf{D}_N\mathbf{C}_N^{-1}\| < \infty$. Then, the proof of Corollary 3 follows from applying standard random matrix arguments to the model of $\hat{\mathbf{S}}_N$ in Theorem 1, considered now as a random matrix in both $\mathbf{y}_i$ and $\mathbf{a}_i$. The result may be straightforwardly obtained from, e.g., [25, Thm. 1] (see Appendix B for similar applications).

Appendix B
Random Matrix Results

In this section we list several intermediary results needed in Appendix A.

Lemma 1. *Let Assumptions 1-2 hold. Define*

$$\mathbf{F}_N \triangleq \frac{1}{n}\sum_{j=1}^{(1-\varepsilon_n)n} v(\gamma_n)\,\mathbf{x}_j\mathbf{x}_j^\dagger + \frac{1}{n}\sum_{j=1}^{\varepsilon_n n} v(\alpha_{j,n})\,\mathbf{a}_j\mathbf{a}_j^\dagger$$

and $\mathbf{F}_{N,(i)} = \mathbf{F}_N - \frac{1}{n}v(\gamma_n)\mathbf{x}_i\mathbf{x}_i^\dagger$*, with* γ_n *and* $\alpha_{j,n}$ *given in Theorem 1. Then, as* $n \to \infty$*,*

$$\max_{1\leq i\leq \varepsilon_n n}\left|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \gamma_n\right| \xrightarrow{\text{a.s.}} 0.$$

Proof: We first need to establish a result on $\lambda_1(\mathbf{F}_{N,(i)})$, for which we know that $\lambda_1(\mathbf{F}_{N,(i)}) \geq \lambda_1(v(\gamma_n)\frac{1}{n}\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j^\dagger)$. Then, [18, Lemma 1] along with Assumption 2 and the boundedness of γ_n show that there exists $\xi > 0$ such that, for all large n a.s.,

$$\min_{1\leq i\leq (1-\varepsilon_n)n} \lambda_1\left(\mathbf{F}_{N,(i)}\right) > \xi. \tag{13}$$

With this acquired, the outline of the proof is divided into two main steps. We first prove that $\max_{1\leq i\leq (1-\varepsilon_n)n}|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}| \xrightarrow{\text{a.s.}} 0$ using quadratic form-close-to-the trace and rank-one perturbation arguments. Then, using [25, Thm 1], we show that $\left|\frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1} - \gamma_n\right| \xrightarrow{\text{a.s.}} 0$.

The triangle inequality allows us to write

$$\left|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\,\mathbf{F}_N^{-1}\right| \leq \left|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\,\mathbf{F}_{N,(i)}^{-1}\right| + \left|\frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-1} - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right|. \tag{14}$$

Let us bound the two terms on the right hand side of (14). Denote by $\mathrm{E}_{\mathbf{x}_i}$ the expectation with respect to $\mathbf{x}_i$ (i.e., conditionally on $\mathbf{F}_{N,(i)}$) and $\kappa_i \triangleq \mathbf{1}_{\{\lambda_1(\mathbf{F}_{N,(i)})>\xi\}}$ with ξ defined in (13). For the first term, we can apply [27, Lemma B.26] (since $\mathbf{x}_i$ is independent of $\kappa_i^{1/p}\mathbf{F}_{N,(i)}^{-1}$), so that for $p \geq 2$,

$$\mathrm{E}_{\mathbf{x}_i}\left[\kappa_i\left|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-1}\right|^p\right] \leq \frac{\kappa_i K_p}{N^{p/2}}\left[\left(\frac{\nu_4}{N}\operatorname{tr}\left(\mathbf{F}_{N,(i)}^{-1}\right)^2\right)^{p/2} + \frac{\nu_{2p}}{N^{p/2}}\operatorname{tr}\mathbf{F}_{N,(i)}^{-p}\right] \tag{15}$$

for some constant K_p depending only on p, with ν_ℓ any value such that $\mathrm{E}\left[|x_{ij}|^\ell\right] \leq \nu_\ell$. Using $\frac{1}{N^k}\operatorname{tr}\mathbf{B}^k \leq \left(\frac{1}{N}\operatorname{tr}\mathbf{B}\right)^k$ for $\mathbf{B} \in \mathbb{C}^{N\times N}$ nonnegative definite and $k \geq 1$ leads to

$$\begin{aligned}\mathrm{E}_{\mathbf{x}_i}\left[\kappa_i\left|\frac{1}{N}\mathbf{x}_i^\dagger\mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-1}\right|^p\right] &\leq \frac{\kappa_i K_p}{N^{p/2}}\left(\nu_4^{p/2} + \nu_{2p}\right)\left(\frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-2}\right)^{p/2} \\ &\leq \frac{K_p}{\xi^p N^{p/2}}\left(\nu_4^{p/2} + \frac{\nu_{2p}}{N^{p/2-1}}\right)\end{aligned} \tag{16}$$

where for the second inequality we have used $\operatorname{tr}\mathbf{B} \leq \|\mathbf{B}\|$ for $\mathbf{B} \in \mathbb{C}^{N\times N}$ nonnegative definite and the fact that $\kappa_i \|\mathbf{F}_{N,(i)}^{-1}\| < \xi^{-1}$, which holds from the definition of κ_i. The bound (16) being irrespective of $\mathbf{F}_{N,(i)}$, we can now take the expectation over $\mathbf{F}_{N,(i)}$ to obtain

$$\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-1}\right|^p\right] = \mathcal{O}\left(\frac{1}{N^{p/2}}\right). \tag{17}$$

For the second term in (14), we can write $\mathbf{F}_{N,(i)} = (\mathbf{F}_{N,(i)} - \frac{\xi}{2}\mathbf{I}_N) + \frac{\xi}{2}\mathbf{I}_N$ with $\mathbf{F}_{N,(i)} - \frac{\xi}{2}\mathbf{I}_N \succ \mathbf{0}$ and we have from [19, Lemma 2.6] (rank-one perturbation lemma)

$$\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\operatorname{tr}\mathbf{F}_{N,(i)}^{-1} - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right|^p\right] \leq \frac{1}{N^p}\left(\frac{2}{\xi}\right)^p. \tag{18}$$

From (14), we can now use Hölder's inequality and the bounds (17)–(18) to obtain

$$\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right|^p\right] = \mathcal{O}\left(\frac{1}{N^{p/2}}\right). \tag{19}$$

Then, we have that

$$\begin{aligned}
&\Pr\left[\max_{1\leq i\leq (1-\varepsilon_n)n} \kappa_i^{1/p}\left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right| > \zeta\right] \\
&\leq \sum_{i=1}^{(1-\varepsilon_n)n} \Pr\left[\kappa_i^{1/p}\left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right| > \zeta\right] \\
&\leq \frac{(1-\varepsilon_n)n}{\zeta^p}\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right|^p\right] \\
&= \mathcal{O}\left(\frac{1}{N^{p/2-1}}\right)
\end{aligned}$$

where we have used (in order) Boole's inequality, Markov's inequality, and (19). Recall from (15) that the entries of $\mathbf{x}_i$ are required to have finite $2p$-th order moment and that, by our initial assumption, $\mathrm{E}[|x_{ij}|^{8+\eta}] < \infty$ for some $\eta > 0$. Then, taking $p > 4$, the Borel Cantelli lemma along with the fact that $\min_{1\leq i\leq (1-\varepsilon_n)n} \kappa_i \xrightarrow{\text{a.s.}} 1$ ensure

$$\max_{1\leq i\leq (1-\varepsilon_n)n}\left|\frac{1}{N}\mathbf{x}_i^\dagger \mathbf{F}_{N,(i)}^{-1}\mathbf{x}_i - \frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}\right| \xrightarrow{\text{a.s.}} 0. \tag{20}$$

It remains to show that γ_n is a deterministic equivalent for $\frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1}$. From (13) and the fact that any subtraction of a nonnegative definite matrix cannot increase the smallest eigenvalue, we have that $\lambda_1(\mathbf{F}_N) > \xi$ for all large n a.s. Then, we can write $\mathbf{F}_N = (\mathbf{F}_N - \frac{\xi}{2}\mathbf{I}_N) + \frac{\xi}{2}\mathbf{I}_N$ with $\liminf_n \lambda_1(\mathbf{F}_N - \frac{\xi}{2}\mathbf{I}_N) > 0$ a.s. and we are in position to apply [25, Thm. 1] which ensures

$$\left|\frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1} - \frac{1}{N}\operatorname{tr}\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+e_N}\mathbf{I}_N + \mathbf{A}_N\right)^{-1}\right| \xrightarrow{\text{a.s.}} 0$$

where $\mathbf{A}_N = \frac{1}{n}\sum_{j=1}^{\varepsilon_n n} v(\alpha_{j,n})\mathbf{a}_j\mathbf{a}_j^\dagger$ and e_N is the unique positive solution to

$$e_N = c_n v(\gamma_n)\frac{1}{N}\operatorname{tr}\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+e_N}\mathbf{I}_N + \mathbf{A}_N\right)^{-1}.$$

According to the definition of γ_n, $e_N = c_n v(\gamma_n)\gamma_n$ with γ_n the solution to

$$\gamma_n = \frac{1}{N}\operatorname{tr}\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+c_n v(\gamma_n)\gamma_n}\mathbf{I}_N + \mathbf{A}_N\right)^{-1}$$

which has been proven to be unique. Altogether,

$$\left|\frac{1}{N}\operatorname{tr}\mathbf{F}_N^{-1} - \gamma_n\right| \xrightarrow{\text{a.s.}} 0. \tag{21}$$

Combining (20) and (21) concludes the proof. ∎

Lemma 2. *Let Assumptions 1-2 hold and define*

$$\mathbf{G}_{N,(i)} \triangleq \frac{1}{n}\sum_{j=1}^{(1-\varepsilon_n)n} v(\gamma_n)\mathbf{x}_j\mathbf{x}_j^\dagger + \frac{1}{n}\sum_{j\neq i} v(\alpha_{j,n})\mathbf{a}_j\mathbf{a}_j^\dagger$$

with γ_n and $\alpha_{j,n}$ defined as in Theorem 1. Then, as $n\to\infty$,

$$\max_{1\leq i\leq \varepsilon_n n}\left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right| \xrightarrow{\text{a.s.}} 0.$$

Proof: Since $\lambda_1(\mathbf{G}_{N,(i)}) \geq \lambda_1(v(\gamma_n)\frac{1}{n}\sum_{j=1}^{(1-\varepsilon_n)n}\mathbf{x}_j\mathbf{x}_j^\dagger)$, we can use [18, Lemma 1] along with Assumption 2 and the uniform boundedness of γ_n to show that there exists $\xi > 0$ such that, for all large n a.s.

$$\min_{1\leq i\leq \varepsilon_n n}\lambda_1\left(\mathbf{G}_{N,(i)}\right) > \xi.$$

Denote $\kappa_i \triangleq \mathbf{1}_{\{\lambda_1(\mathbf{G}_{N,(i)})>\xi\}}$. Using similar derivations as for [28, Lemma 3] adapted to the present model, we have

$$\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right|^p\right] = \mathcal{O}\left(\frac{1}{N^{p/2}}\right). \tag{22}$$

Then

$$\begin{aligned}
&\Pr\left[\max_{1\leq i\leq \varepsilon_n n}\kappa_i^{1/p}\left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right| > \zeta\right] \\
&\leq \sum_{i=1}^{\varepsilon_n n}\Pr\left[\kappa_i^{1/p}\left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right| > \zeta\right] \\
&\leq \frac{\varepsilon_n n}{\zeta^p}\mathrm{E}\left[\kappa_i \left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right|^p\right] \\
&= \mathcal{O}\left(\frac{1}{N^{p/2-1}}\right)
\end{aligned}$$

where we used (in order) Boole's inequality, Markov's inequality, and (22). Taking $p > 4$, the Borel Cantelli lemma ensures

$$\max_{1\leq i\leq \varepsilon_n n}\kappa_i^{1/p}\left|\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{G}_{N,(i)}^{-1}\mathbf{a}_i - \alpha_{i,n}\right| \xrightarrow{\text{a.s.}} 0$$

which then proves Lemma 2 using $\min_{1\leq i\leq \varepsilon_n n}\kappa_i \xrightarrow{\text{a.s.}} 1$. ∎

Appendix C
Asymptotic moments

In this last appendix, we derive the moments of the deterministic equivalents studied in [25]. We provide in full the generic result, which may be used for independent purposes. We first recall [25, Thm. 1].

Theorem 2 (Wagner et al., [25])**.** *Let $\mathbf{Y} \in \mathbb{C}^{N\times n}$ have independent columns $\mathbf{y}_i = \mathbf{H}_i\mathbf{x}_i$, where $\mathbf{x}_i \in \mathbb{C}^{N_i}$ has*

i.i.d. entries of zero mean, variance $1/n$, and $4+\eta$ moment of order $\mathcal{O}(1/n^{2+\eta/2})$, and $\mathbf{H}_i \in \mathbb{C}^{N\times N_i}$ such that $\mathbf{R}_i \triangleq \mathbf{H}_i\mathbf{H}_i^\dagger$ has uniformly bounded spectral norm over n, N. Let also $\mathbf{A}_N \in \mathbb{C}^{N\times N}$ be Hermitian non-negative and denote $\mathbf{F}_N = \mathbf{Y}\mathbf{Y}^\dagger + \mathbf{A}_N$. Then, as N, $N_1, \ldots, N_n$, and n grow large with ratios $c_i = N_i/n$, and $c_0 = N/n$ satisfying $0 < \liminf_n c_i \leq \limsup_n c_i < \infty$ for $0 \leq i \leq n$, we have

$$\frac{1}{n}\operatorname{tr}\left(\mathbf{F}_N - z\mathbf{I}_N\right)^{-1} - m_N(z) \xrightarrow{\text{a.s.}} 0$$

with

$$m_N(z) = \frac{1}{n}\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+e_{N,i}(z)}\mathbf{R}_i + \mathbf{A}_N - z\mathbf{I}_N\right)^{-1} \tag{23}$$

where $e_{N,1}(z), \ldots, e_{N,n}(z)$ form the unique solution of

$$e_{N,j}(z) = \frac{1}{n}\operatorname{tr}\mathbf{R}_j\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+e_{N,i}(z)}\mathbf{R}_i + \mathbf{A}_N - z\mathbf{I}_N\right)^{-1}$$

such that all $e_{N,j}(z)$ are Stieltjes transforms of a non-negative finite measure on $\mathbb{R}^+$.

From Theorem 2, the distribution function F_N with Stieltjes transform $m_N(z)$ is a deterministic equivalent for the eigenvalue distribution of $\mathbf{F}_N$. We next describe the successive moments of the distribution function F_N. This generalizes the asymptotic moment results in [29], valid only for $\mathbf{A}_N = \mathbf{0}$.

Theorem 3. *Let F_N be the distribution function associated with the Stieltjes transform (23), and denote $M_{N,0}, M_{N,1}, \ldots$ the successive moments of F_N, i.e., $M_{N,p} \triangleq \int x^p dF_N$. Then,*

$$M_{N,p} = \frac{(-1)^p}{p!}\frac{1}{N}\operatorname{tr}\mathbf{T}_p$$

with $\mathbf{T}_0, \mathbf{T}_1, \ldots$ defined recursively from

$$\mathbf{T}_{p+1} = -\sum_{i=0}^{p}\mathbf{T}_{p-i}\mathbf{A}_N\mathbf{T}_i + \sum_{i=0}^{p}\sum_{j=0}^{i}\binom{p}{i}\binom{i}{j}\mathbf{T}_{p-i}\mathbf{Q}_{i-j+1}\mathbf{T}_j$$

$$\mathbf{Q}_{p+1} = \frac{p+1}{n}\sum_{k=1}^{n} f_{k,p}\mathbf{R}_k$$

$$f_{k,p+1} = \sum_{i=0}^{p}\sum_{j=0}^{i}\binom{p}{i}\binom{i}{j}(p-i+1)f_{k,j}f_{k,i-j}\beta_{k,p-i}$$

$$\beta_{k,p+1} = \frac{1}{n}\operatorname{tr}\left[\mathbf{R}_k\mathbf{T}_{p+1}\right]$$

and $\mathbf{T}_0 = \mathbf{I}_N$, $f_{k,0} = -1$, $\beta_{k,0} = \frac{1}{n}\operatorname{tr}\mathbf{R}_k$ for $k \in \{1, \ldots, n\}$.

Proof: Follows the same steps as the proof of [29, Thm. 2] with proper modifications to account for $\mathbf{A}_N \neq \mathbf{0}$. ∎

REFERENCES

[1] P. Bianchi, J. Najim, M. Maida, and M. Debbah, "Performance analysis of some eigen-based hypothesis tests for collaborative sensing," in *IEEE Stat. Signal Process. (SSP '09)*, Cardiff (UK), Sep. 2009, pp. 5–8.

[2] B. Nadler, "Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2746–2756, 2010.

[3] X. Mestre and M. A. Lagunas, "Modified subspace algorithms for DoA estimation with large arrays," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 598–614, 2008.

[4] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Stat.*, vol. 36, no. 1, pp. 199–227, 2008.

[5] N. El Karoui, "The spectrum of kernel random matrices," *Ann. Stat.*, vol. 38, no. 1, pp. 1–50, 2010.

[6] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Stat.*, vol. 35, no. 1, pp. 73–101, 1964.

[7] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Stat.*, vol. 4, no. 1, pp. 51–67, 1976. [Online]. Available: http://dx.doi.org/10.1214/aos/1176343347

[8] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Stat.*, pp. 234–251, 1987.

[9] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of Maronna's M-estimator with elliptically distributed samples," *arXiv preprint arXiv:1311.7034*, 2013.

[10] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Multivar. Anal.*, vol. 131, pp. 99–120, 2014.

[11] R. Couillet, A. Kammoun, and F. Pascal, "Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals," *(submitted to) J. Multivar. Anal.*, 2014. [Online]. Available: http://arxiv.org/abs/1410.0817

[12] T. Zhang, X. Cheng, and A. Singer, "Marchenko-Pastur law for Tyler's and Maronna's M-estimators," *http://arxiv.org/abs/1401.3424*, 2014.

[13] L. Yang, R. Couillet, and M. McKay, "Minimum variance portfolio optimization with robust shrinkage covariance estimation," in *IEEE Asilomar Conf. Sig. Sys. Comput.*, Pacific Grove, CA, USA, Nov. 2014.

[14] R. Couillet, "Robust spiked random matrices and a robust G-MUSIC estimator," *(submitted to) J. Multivar. Anal.*, 2014. [Online]. Available: http://arxiv.org/pdf/1404.7685

[15] Y. Chitour, R. Couillet, and F. Pascal, "On the convergence of maronna's m-estimators of scatter," *IEEE Signal Process. Lett.*, vol. 22, no. 6, pp. 709–712, 2014.

[16] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *Ann. Stat.*, pp. 2102–2119, 1991.

[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[18] R. Couillet, F. Pascal, and J. Silverstein, "Robust estimates of covariance matrices in the large dimensional regime," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7269–7278, 2014.

[19] J. W. Silverstein and Z. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *J. Multivar. Anal.*, vol. 54, no. 2, pp. 175–192, 1995.

[20] L. Laloux, P. Cizeau, M. Potters, and J. P. Bouchaud, "Random matrix theory and financial correlations," *Int. J. Theoretical Appl. Finance*, vol. 3, no. 3, pp. 391–397, Jul. 2000.

[21] A. A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I.-M. Hsing, and M. R. McKay, "Statistical linkage of mutations in the non-structural proteins of hepatitis C virus exposes targets for immunogen design," *J. Virology*, vol. 88, no. 13, pp. 7628–7644, 2014.

[22] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4097–4107, 2011.

[23] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[24] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, 1995.

[25] S. Wagner, R. Couillet, M. Debbah, and D. T. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, 2012.

[26] Z. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices," *Ann. Prob.*, pp. 316–345, 1998.

[27] Z. D. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. New York, NY, USA: Springer Series in Statistics, 2009.

[28] W. Hachem, P. Loubaton, X. Mestre, J. Najim, and P. Vallet, "A subspace estimator for fixed rank perturbations of large random matrices," *J. Multivar. Anal.*, vol. 114, pp. 427–447, 2013.

[29] J. Hoydis, M. Debbah, and M. Kobayashi, "Asymptotic moments for interference mitigation in correlated fading channels," in *IEEE Int. Symp. Inf. Theory (ISIT)*, St. Petersburg (Russia), Jul. 2011, pp. 2796–2800.